Bioinformatics of Eukaryotic Gene Regulation

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr Dipl.-Phys. Szymon M. Kielbasa
geboren am 12.03.1973 in Krakau, Polen

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Thomas Buckhout, PhD

Gutachter:

1. Prof. Dr. Hanspeter Herzel
2. Prof. Dr. Joachim Selbig
3. Prof. Dr. Martin Vingron

Tag der mündlichen Prüfung: 27. Februar 2006

Abstract

Understanding the mechanisms which control gene expression is one of the fundamental problems of molecular biology. Detailed experimental studies of regulation are laborious due to the complex and combinatorial nature of interactions among involved molecules. Therefore, computational techniques are used to suggest candidate mechanisms for further investigation.

This thesis presents three methods improving the predictions of regulation of gene transcription. The first approach finds binding sites recognized by a transcription factor based on statistical over-representation of short motifs in a set of promoter sequences. A succesful application of this method to several gene families of yeast *Saccharomyces cerevisiae* is shown. More advanced techniques are needed for the analysis of gene regulation in higher eukaryotes. Hundreds of profiles recognized by transcription factors are provided by libraries. Dependencies between them result in multiple predictions of the same binding sites which need later to be filtered out. Therefore, the second method presented here offers a way to reduce the number of profiles by identifying similarities between them. Still, the complex nature of interaction between transcription factors makes reliable predictions of binding sites difficult. Exploiting independent sources of information reduces the false predictions rate. The third method described here proposes a novel approach associating gene annotations with regulation of multiple transcription factors and binding sites recognized by them. The utility of the method is demonstrated on several well-known sets of transcription factors.

Although the regulation of transcription is the major cellular mechanism of controlling gene expression, RNA interference provides a way of efficient down-regulation of specific genes in experiments. Difficulties in predicting efficient siRNA sequences motivated the development of a library containing siRNA sequences and related experimental details described in the literature. This library, presented in details in the last chapter, is publicly available at `http://www.human-sirna-database.net`.

Keywords:
prediction of transcription factor binding sites, prediction of transcription factors functions, regulation of gene expression, similarity of transcription factor profiles

Zusammenfassung

Die Aufklärung der Mechanismen zur Kontrolle der Genexpression ist eines der wichtigsten Probleme der modernen Molekularbiologie. Detaillierte experimentelle Untersuchungen sind enorm aufwändig aufgrund der komplexen und kombinatorischen Wechselbeziehungen der beteiligten Moleküle. Infolgedessen sind bioinformatische Methoden unverzichtbar bei der Suche nach neuen Hypothesen, die dann in den Experimenten überprüft werden können. Diese Dissertation stellt drei Methoden vor, die die Vorhersage der regulatorischen Elementen der Gentranskription verbessern. Der erste Ansatz findet Bindungsstellen, die von den Transkriptionsfaktoren erkannt werden. Es basiert auf der statistischen Überrepräsentation von kurzen Motiven in einer Menge von Promotersequenzen. Eine erfolgreiche Anwendung dieser Methode in der Hefe *Saccharomyces cerevisiae* wird vorgestellt.

Weiter fortgeschrittene Techniken sind allerdings notwendig, um die Genregulation in höheren Eukaryoten zu analysieren. In verschiedenen Datenbanken liegen Hunderte von Profilen vor, die von den Transkriptionsfaktoren erkannt werden. Die Ähnlichkeit zwischen ihnen resultiert in mehrfachen Vorhersagen einer einzigen Bindestelle, was im Nachhinein korrigiert werden muss. Es wird deswegen eine Methode vorgestellt, die eine Möglichkeit zur Reduktion der Anzahl von Profilen bietet, indem sie die Ähnlichkeiten zwischen ihnen identifiziert. Die komplexe Natur der Wechselbeziehung zwischen den Transkriptionsfaktoren macht jedoch die Vorhersage von Bindestellen schwierig.

Auch mit einer Verringerung der zu suchenden Profile sind die Resultate der Vorhersagen noch immer stark fehlerbehafted. Die Zuhilfenahme der unabhängigen Informationsressourcen reduziert die Häufigkeit der Falschprognosen. Die dritte beschriebene hier Methode schlägt einen neuen Ansatz vor, die die Gen-Anotation mit der Regulierung von multiplen Transkriptionsfaktoren und den von ihnen erkannten Bindestellen assoziiert. Der Nutzen dieser Methode ist demonstriert am Beispiel von verschiedenen wohlbekannten Sätzen von Transkriptionsfaktoren.

Obwohl die Regulation der Transkription der wichtigste Mechanismus zur Kontrolle der Genexpression ist, bietet die RNA-Interferenz einen effizienten experimentellen Weg zur gezielten Genausschaltung. Die Schwierigkeiten in der Vorhersage von effizienten siRNA Sequenzen motivierte die Entstehung einer Bibliothek mit solchen Sequenzen und dazugehörigen experimentellen Details, die der Literatur entnommen sind. Die Bibliothek,

beschrieben im letzten Kapitel, ist öffentlich zugänglich unter `http://www.human-sirna-database.net`.

Schlagwörter:
Vorhersage von Transkriptionsfaktor-Bindungsstellen, Vorhersage der Funktion von Transkriptionsfaktor, Regulation von Gen-Expression, Änlichkeit von Transkriptionsfaktor-Profilen

Contents

1 Introduction **1**

2 Overrepresented words as regulatory elements **7**
2.1 Introduction 8
2.2 Materials and Methods 9
2.2.1 Scoring motif frequencies 10
2.2.2 Scoring positional information 11
2.3 Results 13
2.3.1 Evaluation of the frequency score 13
2.3.2 Frequency score: variation of the parameters 16
2.3.3 Combination of both scores 17
2.4 Discussion 20

3 Regulatory elements of AP-1 and RAS targets **23**
3.1 Introduction 24
3.2 Study of AP-1 regulated genes 25
3.3 Study of the RAS-dependent genes 28
3.4 Discussion 30

4 Similarities of profiles recognized by transcription factors **35**
4.1 Introduction 36
4.2 Methods 37
4.2.1 Jaspar and Transfac databases 37
4.2.2 χ^2-based distance D between position frequency matrices 39
4.2.3 Correlation C of position weight matrices scores . . . 41
4.3 Results 43
4.3.1 Comparison of both similarity measures 43
4.3.2 Clusters of similar matrices in Jaspar and Transfac databases 46
4.3.3 Mapping of novel matrices 49

4.4 Discussion 50

5 Prediction of functions of transcription factors (TFGossip) 53
5.1 Introduction 54
5.2 Materials and Methods 56
5.2.1 TFGossip 56
5.2.2 Gossip 56
5.2.3 Data preparation 58
5.3 Results 59
5.3.1 Functions of E2F transcription factor 59
5.3.2 Functions of NFAT and AP-1 transcription factors . . . 59
5.3.3 Processes regulated by muscle transcription factors . . 61
5.4 Discussion 61

6 Human siRNA Database (HuSiDa) 69
6.1 Introduction 70
6.2 Results 71
6.2.1 HuSiDa – database 71
6.2.2 HuSiDa – web interface 72
6.3 Discussion 73

7 Outlook 77

A Overrepresented words as regulatory elements 81
A.1 Z-score formula 81

B Lists of similar profiles 85

List of Figures

2.1 Overview of Z-score calculation 11
2.2 Known binding sites of yeast gene families 12
2.3 A pattern found in promoters of H-Ras regulated human genes 15
2.4 Frequency and positional scores in several families. 19

3.1 Predictions of AP-1 binding sites 27
3.2 Predictions in promoters of RAS target genes 29
3.3 Distance of binding sites as a function of a score threshold . . 31
3.4 GC-content of the promoters the RAS-dependent genes. 32

4.1 Variables of χ^2-based similarity definition 39
4.2 Example of χ^2 based distance: CREB and ATF matrices . . . 40
4.3 Weight matrix score calculation 43
4.4 Matrices of CREB and ATF: correlation of scores 44
4.5 Comparison of similarity measures of matrices 45
4.6 Similar Jaspar and Transfac matrices 47
4.7 Examples of clusters of similar matrices 48
4.8 Mapping of CLOCK-BMAL1 matrices 49

5.1 Data flow in the TFGossip algorithm 56
5.2 Details of the TFGossip algorithm 57
5.3 Predicted functions of E2F transcription factors 60
5.4 Predicted functions of NFAT and AP-1 transcription factors . 65
5.5 Skeletal muscle genes transcription factors (details) 66
5.6 Skeletal muscle genes transcription factors (context) 67

6.1 Human siRNA Database search form 73
6.2 A browser of records matching a database query 74
6.3 A view of a siRNA record 75

List of Tables

2.1 Known and predicted motifs in yeast families 14
2.2 Top overrepresented words in two human families 15
2.3 Overrepresented motifs as a function of the motif length . . . 16
2.4 Overrepresented motifs for different background models 17
2.5 Overrepresented motifs for different upstream lengths 18

3.1 Predictions of AP-1 binding sites 26

4.1 Properties of Jaspar and Transfac matrices 38
4.2 Differences of nucleotide distributions reported by $\chi 2$ measure 40

B.1 Matrices of Jaspar mapped to Transfac 86
B.2 Clusters of matrices in Jaspar and Transfac 89

Chapter 1

Introduction

Eukaryotes, from yeast to man, maintain diverse sets of genes whose expression levels are modulated to satisfy the demands of developmental, environmental or physiological conditions. Although the abundance of proteins can be controlled through a variety of mechanisms, alteration of gene transcriptional rates is the most direct and utilized cellular tool [Wasserman and Fickett, 1998]. Transcription factors influence gene expression as an effect of binding to regulatory sites typically located in promoters or enhancers of the corresponding genes. Therefore, understanding the language of binding sites describing the influence of transcription factors on the expression of corresponding genes is one of the fundamental challenges of molecular biology.

Effects of binding of single or several transcription factors in a promoter of a specified gene may be studied individually in laborious experiments. Complex interactions of many involved components may be analysed (e.g. the analysis of the sea urchin Endo 16 gene showed that the upstream region of this gene contains at least 33 transcription factors binding sites in five modules, Yuh et al. [1998]).

Results of such experiments corresponding to particular transcription factors may be collected together in order to construct a model of the sites recognized by the factor (an example is given in chapter 3, where I construct a set of sites recognized by AP-1). To ensure maximum specificity of such a model only those binding sites are desired for which there exist clear and direct evidence both for function and identity of the transcription factor bound. Analysis of such DNA binding sites can be conveniently divided into two subproblems [Stormo, 2000]. The first is, given a collection of known binding sites of a single transcription factor, to develop a representation of those sites which can be used to search new sequences and reliably predict where additional binding sites occur. The choice of the representation depends on the number of available known binding sites and how precisely the

experimentally found binding motifs have been identified. The second problem is, given a set of DNA sequences expected to contain binding sites for a common transcription factor, but not knowing where the sites are, to discover the location of the sites in each sequence. Here, a conceptually simple procedure – scoring binding site models along the sequences results in a *curse of false positives*. The rate of transcription factor binding site predictions varies for different binding models and their parameters, but typically a candidate site is reported every 500-5000 bp [Wasserman and Sandelin, 2004]. It has been shown that employing other features of binding sites helps to reduce the number of false predictions by an order of magnitude.

Roulet et al. [2002] propose a computationally driven method combining systematic evolution of ligands by exponential enrichment (SELEX) and serial analysis of gene expression (SAGE, Velculescu et al. [1995]). Their technique allowed the authors to construct a high-quality model of interaction between DNA and the CTF/NFI transcription factor. Moreover, whole chromosome studies utilizing chromatin immunoprecipitation and genomic microarray techniques (ChIP on chip) provide a method for measuring locations of a large number of transcription factor binding sites simultaneously. For example, the sites bound by p65 (belonging to NF-κB family) [Martone et al., 2003] and by CREB [Euskirchen et al., 2004] were globally identified on the human chromosome 22. The results showed, that binding was not restricted to promoter regions; the sites were found elsewhere, including introns and unannotated chromosome regions. Moreover, binding was observed in front of genes whose expression was not altered, thereby suggesting that binding alone was not sufficient for gene activation. This observation motivates the development of concepts taking into account cooperative interactions between transcription factors, like the method for inferring functions of transcription factors (which will be presented in chapter 5).

Sequencing of the yeast *Saccharomyces cerevisiae* genome as well as large scale gene expression studies provided data motivating the design of *ab-initio* regulatory elements prediction algorithms. Analysis of the levels of gene expression after a known stimuli allows to extract clusters of genes following the same behavior in several time points or experimental conditions. In such cases co-expressed genes are assumed to be co-regulated by the externally induced factor, although it can not be excluded that the co-expression occurs by mere coincidence [Pilpel et al., 2001].

Several methods have been developed to identify the corresponding binding sites, provided that short ($<$ 2 kb) promoter sequences containing the sites can be reliably extracted. Since the non-coding parts of the yeast genome are relatively short, and using of the upstream untranslated se-

quences is a good approximation of the promoter regions, several algorithms successfully predict the right sites:

- Gibbs sampler [Lawrence et al., 1993], AlignACE [Roth et al., 1998] detect motifs by aligning short fragments of the input promoter sequences, based on the statistical method of iterative sampling. GLAM [Frith et al., 2004b] extends this approach by providing a way to determine the width of the aligned motif and to calculate the statistical significance of the alignment. MEME [Bailey and Elkan, 1994] uses expectation maximization and artificial intelligence heuristics to construct an alignment. In all cases the outcome is a matrix describing frequencies of nucleotides at the motif positions (PFM, positional frequency matrix). Its quality is growing with the number of aligned sequences.

- When the number of aligned sequences is small, providing a consensus sequence enumerating nucleotides at the motif positions provides similar information to PFM. Therefore, van Helden et al. [1998] extract regulatory sites from the sequences based on computational analysis of short words frequencies. The words are constructed from the nucleotides `A`, `C`, `G`, and `T`.

- The ITB algorithm [Kielbasa et al., 2001], which will be presented in chapter 2, fills the gap between the above methods. The idea behind our approach bases on the observation of the experimental cases in which the number of available promoter sequences is still small, but the variability of the binding sites is such, that an extended alphabet containing symbols alternatively representing multiple nucleotides are more appropriate.

The methods presented above were sufficient to detect a large class of regulatory elements in yeast but, as demonstrated in chapter 3, their applicability to human genomic sequences is limited, so more advanced approaches are necessary. Several features of higher eukaryotic sequences are responsible for this complication. The regulatory element prediction algorithms require at their input short sequences expected to contain the binding sites. But detailed experimental studies of promoter locations have been performed for a small fraction of human genes only [Perier et al., 1998]. Therefore, promoter prediction methods [Davuluri et al., 2001, Ohler et al., 2001, Liu and States, 2002, Scherf et al., 2000] need to be applied to the gene upstream sequences, which ideally should handle the issue of long introns, non-coding

first exons and multiple transcripts. Moreover, growing complexity of regulatory interactions makes it difficult to distinguish direct and indirect effects in high throughput gene expression studies. This way the chances grow, that a gene is classified as regulated by a certain factor although it rather belongs to a cascade induced by the factor. Moreover, examples show, that there are cases where regulatory elements are located far from the regulated gene ($\approx$ 40 kb) [Gottgens et al., 2000]. Due to these reasons a binding site prediction algorithm needs to take into account that some of supplied sequences have poor overlap with the core promoters.

Consequently, concepts utilizing various properties of regulatory mechanisms have been developed to improve specificity of the predictions:

- Bussemaker et al. [2001], Caselle et al. [2002] propose to eliminate the clustering step needed to create a list of genes believed to be co-regulated. Instead, correlating presence of binding sites with gene expression levels directly is suggested.

- Wagner [1997] proposes a method which takes advantage of the fact that many transcription factors show cooperativity in transcriptional activation. The algorithm detects closely spaced binding sites of the same transcription factor.

- Pilpel et al. [2001], Frith et al. [2002, 2003], Murakami et al. [2004] construct techniques scoring overrepresented close occurrences of binding sites recognized by different transcription factors. Experimental observations of such pairing are also collected in dedicated databases [Kel-Margoulis et al., 2000]. Furthermore, a computational analysis of the whole human genome is available [Hannenhalli and Levy, 2002].

- Phylogenetic footprinting – preferential conservation of functional sequences over the course of evolution by selective pressure results in a striking enrichment of regulatory sites among the conserved regions [Dieterich et al., 2002, Wasserman et al., 2000]. These lines were followed to combine the knowledge of co-regulation among different genes and conservation among orthologous genes to improve the identification of motifs [Wang and Stormo, 2003, Lenhard et al., 2003].

- Frith et al. [2004a] contributes with a method taking into account in probability calculations the possibility, that a part of genes is incorrectly assigned to co-regulated set of genes.

The majority of the methods presented above require a collection of profiles recognized by transcription factors as the input data. Typically the

provided profiles are treated as mutually independent (i.e. associated with different transcription factors). This assumption is not easy to guarantee, especially if profiles from a library (Jaspar [Sandelin et al., 2004a], Transfac [Wingender et al., 1996, 2000, Matys et al., 2003]) are used. Moreover, in some applications the libraries providing several hundred of records are too large and selection of a core subset is needed. These observations motivated the development of measures allowing to compare and filter the profiles recognized by transcription factors. The details of this study are provided in chapter 4.

It has been widely accepted that improving the quality of transcription factor binding site predictions requires to employ many of the binding site properties in a statistically correct way [Sandelin et al., 2004b]. Since each of these features may reduce the number of the false positive predictions, following Kielbasa et al. [2004a] and Blüthgen et al. [2005b] I present in chapter 5 a method associating cooperative binding of transcription factors with biological functions of the corresponding genes. This novel approach, utilizing growing public gene annotations, not only provides a new technique for narrowing the list of genes involved in regulation of a process, but it also allows direct inferring the processes controlled by studied factors.

Finally it should be mentioned, that practical applications of the presented tools require comfortable and easy user interfaces. The development of HomGL [Blüthgen et al., 2004], a web-based application for retrieval of upstream homologous human/mouse/rat sequences, was partially influenced by the requirements of transcription factor binding site prediction methods. Additionally, the SeqVISTA application [Hu et al., 2004], provides a unified interface able to request calculations as well as visualize the results of many algorithms presented above.

Despite the fact, that alteration of gene transcription is the most direct and utilized regulatory tool, in recent years another cellular mechanism attracted efforts of experimental groups. RNA interference (RNAi) is a sequence-specific posttranscriptional mechanism, which is triggered by double-stranded RNA and causes degradation of mRNAs homologous in sequence to the introduced dsRNA Elbashir et al. [2002]. The hallmark of RNAi is its specificity – carefully designed short interfering RNA sequence (siRNA) is able to reduce the expression of the gene from which the sequence is derived, with minor effects on the expression of genes unrelated in sequence [Fire et al., 1998]. Moreover, the induced gene silencing is reversible and thus does not appear to reflect a genetic change. Therefore, siRNAs have provided a new tool for studying gene functions.

It has been observed, that siRNA sequences which target different regions

of the same mRNA vary significantly in their effectiveness [Holen et al., 2002, Reynolds et al., 2004]. Many factors are likely to be responsible for that fact, including nucleotide composition of the siRNA sequence, presence of specific nucleotide patterns, mRNA secondary structure, etc. Therefore, costly search for new siRNAs prompted many groups to design algorithms predicting active sequences for genes specified by a user [Ui-Tei et al., 2004, Amarzguioui and Prydz, 2004, Reynolds et al., 2004, Khvorova et al., 2003, Saetrom, 2004]. Although the quality of predictions is growing, as it has been pointed out by Saetrom and Snove [2004], an independent and publicly available database containing a large collection of verified siRNA sequences should be established. This suggestion motivated us [Truss et al., 2005] to design the Human siRNA Database of the siRNA molecules and important technical details of the corresponding gene silencing experiments which I present in details in chapter 6. The database is available at the address `http://www.human-siRNA-database.net`.

Chapter 2

Overrepresented words as regulatory elements

Summary

Microarray studies analyse expression of a large number of genes in a number of conditions or time points. Groups of genes following similar expression patterns are a typical result of such experiments. Emergence of such a pattern suggests existence of a regulatory mechanism shared within a group. In higher organisms control of gene expression is typically mediated by complex interactions involving many factors (studied later in chapter 5). In contrast, in lower eukaryotes the regulation can often be understood as an effect of a single factor. Assuming an existence of a single transcription factor regulating a given gene family, it is possible to construct an algorithm to discover an unknown binding profile of this factor. This chapter presents a program ITB (Integrated Tool for Box finding, Kielbasa et al. [2001]), capable to build a ranked list of candidate profiles, which are significantly overrepresented in the upstream regions of a group in comparison to a training (reference) set of upstream gene sequences. The profiles are modeled as short words built up of simple nucleotides `A`, `C`, `G` or `T`, as well as their mixed forms (expressed with the IUB nucleotide code). Given a profile length, possible words are exhaustively enumerated while their probabilities are estimated and ranked for final evaluation (discussed in details in section 2.2). The results section 2.3 presents an application of the method to several yeast *Saccharomyces cerevisiae* gene groups and compares them to known experimental data. Chapter 3 discusses usage of the method for higher eukaryotes.

2.1 Introduction

Even though a number of genome projects have been completed on the sequence level, only a small fraction of mechanisms governing gene expression have been identified. Using hundreds of profiles recognized by transcription factors, available in databases (Jaspar [Sandelin et al., 2004a], Transfac [Wingender et al., 1996, 2000, Matys et al., 2003]), a promoter of a gene might be scanned for candidate binding sites [Stormo, 2000]. In practical applications such a simple approach leads to results difficult to interpret. Since experimental locations of real promoters are rarely known, they have either to be computationally predicted or sequences located upstream of gene coding regions are taken instead. In addition, simple techniques of calculating binding probabilities result in a large number of false positive signals [Wasserman and Sandelin, 2004].

Therefore incorporation of independent experimental results may improve the prediction quality. High throughput differential expression measurements provide data describing transcription levels of genes of various model organisms in response to environmental stimuli. Typically, groups of genes following similar expression patterns are constructed with the help of clustering algorithms [Herzel et al., 2001]. Genes belonging to such a group are assumed to be members of a family of genes regulated by the same transcription factor. Based on this assumption, the gene upstream regions are studied in order to extract transcription factor binding sites [Brazma et al., 1998, Roth et al., 1998, van Helden et al., 1998, Zhang, 1999, Hughes et al., 2000].

Several different approaches are used to detect regulatory elements in the promoter/upstream regions of co-regulated genes. An exhaustive algorithm RSA-tools-oligo-analysis [van Helden et al., 1998] compares the frequencies of conserved words (built of nucleotides `A`, `C`, `G` or `T`) in a given set of promoter sequences to the corresponding frequencies in a reference set (in the following termed training set). This method is sensitive in detecting over-represented words in the upstream regions of co-regulated yeast *Saccharomyces cerevisiae* genes. Unfortunately, regulatory elements lacking a conserved core sequence might remain undetected by this method.

Weight matrix based methods like AlignACE [Roth et al., 1998] (based on the Gibbs sampling algorithm, first described by Lawrence et al. [1993]), MEME [Bailey and Elkan, 1994], or recently developed GLAM [Frith et al., 2004b] use a multiple alignment strategy to detect the signals of DNA regulatory elements. In this way, elements lacking a conserved core may be found. However, signals due to regulatory elements involved in transcription are rather weak. If only a small number of motifs occur in the co-regulated

sequence set (or simply a small number of sequences is available), weight matrices are of limited use. In such a case, mono- or dimeric repeats or non-specific signals like the motif `AAATAA` are more likely to be aligned than functional regulatory elements.

An intermediate solution, usable for small sets of co-regulated genes, is to search exhaustively for regulatory elements expressed with symbols matching multiple bases. Based on Kielbasa et al. [2001], this chapter presents such a method – the ITB algorithm (Integrated Tool for Box finding), which integrates frequency and positional information to predict transcription factor binding sites in upstream regions of co-regulated genes. Short regular expression-like patterns, allowing small gaps and matching of more than one nucleotide at a position, are analyzed exhaustively and ranked according to two independent scores. The first one, discussed in details in section 2.2.1, scores overrepresentation of a motif within the co-regulated genes as compared to a training set of genes. An idea similar in spirit was proposed independently by Sinha and Tompa [2000]. The second score, mentioned in section 2.2.2, uses a method developed with Jan Korbel, which ranks preferences of the motif candidates to cluster in a certain location relative to the transcription start sites. Results of applying both of the scores to several yeast *Saccharomyces cerevisiae* families (extracted as described in section 2.2) are presented in details in section 2.3. Additionally, an application of the program to a set of genes predicted to be RAS targets is shown there as well.

2.2 Materials and Methods

Prior to the development of the ITB algorithm a study has been done to identify sets of genes in the yeast *Saccharomyces cerevisiae* genome, known to be regulated by the same transcription factor. This way 11 families of genes were constructed, in which the factors as well as the corresponding binding site profiles were known. Gene families for the transcription factors: Gln3, Cbf1/Met4/Met28, Met31/Met32, Pdr1/Pdr3, Ino2/Opi1, Mig1, Yap1 and Gal4 have been provided by van Helden et al. [1998]. Additionally, a family corresponding to the Matα2 transcription factor was constructed based on DNA microarray experiments that revealed genes with a mating type specific transcriptional regulation [Roth et al., 1998]. The top ten transcripts whose expression levels increased the most in mating type "a" relative to "α" were chosen (genes: MFA1, MFA2, AGA2, STE2, BAR1, PMP1, SRA3, VPS13, YKR071C and YBR147W). DNA sequences were retrieved from the MIPS database [Mewes et al., 1997] by extracting 800 bp long regions upstream of

the corresponding ORFs. As the training set all yeast upstream sequences were used.

Besides, two human sequence sets were studied. Zuber et al. [2000] reported the down-regulation of several genes via the human H-Ras protein involving the RAF/MEK/ERK cascade of cytoplasmic kinases. Five promoter regions of co-regulated genes (LOX, LOXL1, LOXL2, RIL/LIM and TSP1) were kindly provided by the group. Moreover, four promoter sequences of genes (EP11114, EP15041, EP36018 and EP37001) up-regulated via the c-Myc protein [Coller et al., 2000] were extracted from the EPD database [Perier et al., 1998]. In case of human analysis 271 human promoter sequences available in the EPD database were used as the training set. This set did not include the studied genes regulated via H-Ras. For the analysis of c-Myc-controlled genes, the four promoters extracted from EPD were removed from the training set.

The ITB program performs an exhaustive search – scores are calculated for all possible n-mers built of the nucleotide alphabet `ACGT` or from the standard IUB nucleotide code `ACGTWSRKYMN` (where degenerated symbols stand for any of the enumerated nucleotides: `W` $\rightarrow$ `AT`, `S` $\rightarrow$ `CG`, `R` $\rightarrow$ `AG`, `K` $\rightarrow$ `GT`, `Y` $\rightarrow$ `CT`, `M` $\rightarrow$ `AC`, `N` $\rightarrow$ `ACGT`). Symbols matching three nucleotides are not studied, since for small numbers of matches the difference between them and `N` can be neglected. After analysis, both scores are visualized on a scatter plot. On request, self-overlapping patterns with a periodicity of one or two (e.g. `AAAAAA` or `AWAWAW`) are removed from the output.

2.2.1 Scoring motif frequencies

The first score used by the ITB algorithm compares observed frequencies of conserved elements in the given genomic sequence set to their expected frequencies, estimated based on a training (background) set (see Fig. 2.1).

The Z-score function $Z(W)$ [van Helden et al., 2000a] is used as the scoring formula for a motif W. It characterizes the difference between the observed number of occurrences $n_{\mathrm{obs}}(W)$ of the motif W (on any strand) and the expected mean number $\mu(W)$, scaled by the expected standard deviation of the motif $\sigma(W)$:

$$Z(W) = \frac{n_{\mathrm{obs}}(W) - \mu(W)}{\sigma(W)}. \tag{2.1}$$

The standard deviation of the motif is estimated based on expressions which take into account the possibility of motif self-overlapping (e.g. `AAAAAA` or `TATATA`) [Pevzner et al., 1989]. These formulae have been adapted for a double-strand analysis and for the degenerated alphabet (see Appendix A

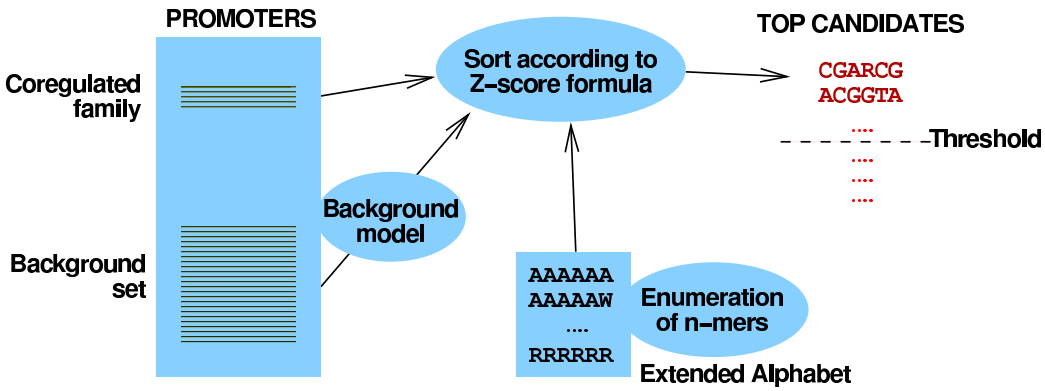


Figure 2.1: Overview of Z-score calculation.

for a detailed derivation):

$$\mu(W) = \sum_{w \in \mathcal{W}} Np(w),$$

$$\sigma^2(W) = \sum_{w \in \mathcal{W}} Np(w)\left(1 - p(w)\right) + 2 \sum_{s=1}^{L_W - 1} (N - sM) \sum_{w,v \in \mathcal{W}} \left(\pi_s^{w,v} - p(w)p(v)\right), \tag{2.2}$$

where N is the total number of possible motif positions in the co-regulated set of M promoters. By definition $\mathcal{W}$ is a set of all oligonucleotides expressed with the simple alphabet ACGT, which match the pattern W or the pattern complementary to it[1]. The number of letters in the pattern W is represented by L_W.

The remaining two symbols $p(w)$ and $\pi_s^{w,v}$ link the Z-score formula to the background model constructed on the training set. $p(w)$ is the estimated probability of the sequence w. $\pi_s^{w,v}$ expresses the probability of a sequence built by overlapping the sequences v at position $s+1$ and w at the first position[2]. These probabilities are estimated with Markov models built on the training set. Here, w and v always denote motifs which consist only the simple nucleotides ACGT.

2.2.2 Scoring positional information

Experimentally verified binding sites of 52 yeast transcription factors were extracted from the SCPD database [Zhu and Zhang, 1999]. While locations

[1]For example, if W = ASTG the described expansion gives $\mathcal{W}$ = (ACTG, AGTG, CAGT, CACT).

[2]For example $\pi_2^{\texttt{ATAC,ACGG}} = p(\texttt{ATACGG})$, $\pi_2^{\texttt{ATAC,TTGG}} = 0$.

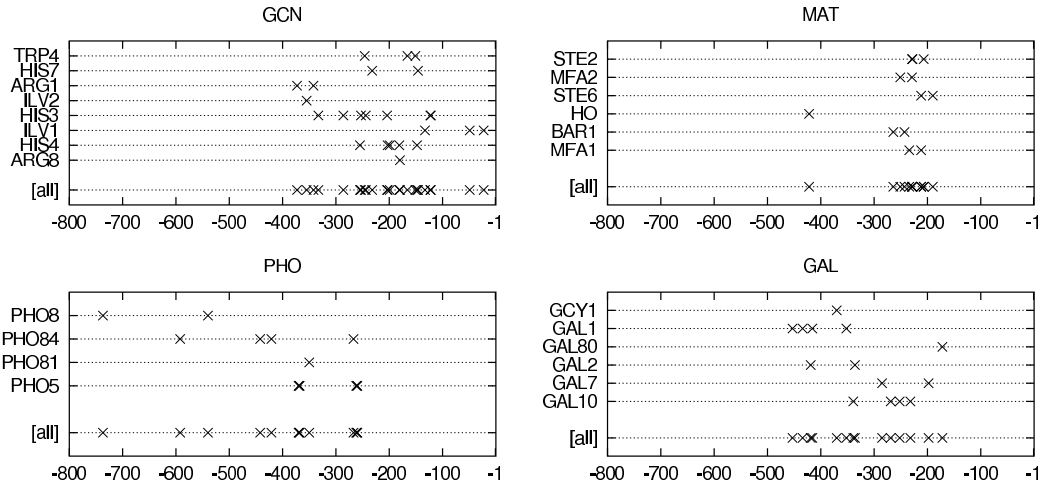


Figure 2.2: The locations of experimentally verified binding sites of Matα2, Gcn4, Pho4, and Gal4 in gene upstream regions. All positions were extracted from SCPD [Zhu and Zhang, 1999] and are relative to the translation start site. The lines at the bottom (denoted "[all]") show positions of all respective transcription factor binding sites marked along a single axis. Gene sets shown here are a part of the regulatory families analyzed in this study.

of some of the sites appear randomly distributed over the corresponding upstream regions, a high proportion of the sites reveals a strong position bias. For Gal4, Gcn4, and Matα2 a clustering was observed, despite the alignment of sequences according to the translation start sites instead of the transcription start sites (see Fig. 2.2). Therefore, the second score used by the ITB algorithm bases on the observation that transcription factor binding sites are often clustered or appear at preferred positions. A procedure (implemented by Jan Korbel in Kielbasa et al. [2001]) scores locations of motifs in gene upstream regions. Positions of all occurrences of a studied motif W within the upstream sequences are collected and marked on a single axis. Then, the lengths of the shortest distances l with a given number of motifs inside ($n = 2, 3, \ldots$) are measured. Probability $p(n, l)$ of observing n motifs over a distance l is estimated with stochastic simulations. The negative logarithm of $p(n, l)$ for the most unprobable pair (n_W, l_W) constitutes the motif positional score.

2.3 Results

2.3.1 Evaluation of the frequency score

The frequency score calculation procedure was applied to the genes belonging to the collected transcription factor families. The simple nucleotide alphabet `ACGT`, as well as the degenerated alphabet `ACGTWRKSYMN` were used. Over-represented motifs of length 6 bp were searched in 800 bp long upstream sequences of the respective genes. The background probabilities were estimated with a Markov model of order 3. An option to remove self-overlapping patterns with a periodicity of 1 or 2 was applied. The motifs of 10 highest Z-scores were studied.

An analysis performed on randomized sets of sequences revealed that the Z-scores were distributed nearly Gaussian, with no more than 0.5% of motifs with the Z-score greater than 3 and with less than 0.0001% of motifs with the Z-score above 5.

Tab. 2.1 shows results of an analysis of 11 co-regulated yeast *Saccharomyces cerevisiae* gene sets. Most previously characterized yeast regulatory elements are correctly predicted. In the families: NIT, MET, INO, PHO, PDR, GCN, YAP, and TUP the highest scoring motifs matched previously characterized transcription factor binding sites. Analysis of the MAT family revealed a motif corresponding to a previously described consensus sequence at rank 3. In the MET family, a second site previously characterized as not directed by the Cbf1/Met4/Met28-complex but by Met31/32 [Kuras et al., 1996] was also found. In most families analyzed, several of the top predictions partly matched previously identified consensus sequences. Principally similar results were obtained independently of the alphabet used.

In addition to matches to previously identified consensus sequences, a number of sequences not matching known transcription factor binding sites were predicted. In particular, the motifs `CAACAA` predicted in the INO family (Z-score 6.9, information content[3] 5.2 bits) and `CGTTCC` (6.3, 8.0 bits) found in the YAP family stand out with significant scores. Several other high scoring motifs, which self-overlap several times, were not considered as promising candidates.

The frequency score analysis failed to detect two known elements – the consensus sequences of the transcription factors regulating the GAL and HAP families.

The publicly available algorithms AlignACE version 3.0 [Roth et al.,

[3]The information content is calculated as in introduced in Schneider et al. [1986] for skewed genomes. Additionally, the sampling error correction by Miller and Quastler [1955] is applied.

Family name (no. of genes), bound factors	Experimentally verified motifs	Predictions: degenerated and normal alphabets (rank)	No. of motifs /seqs. with match	Z-score (Information content [bits])
NIT(7) Gln3	GATAAG[b]	GATAAG(1) GATAAG(1)	26/6	13.9 (7.2)
MET_1(11) Cbf1/Met4/Met28	TCACGTG[b]	CACGTG(1) CACGTG(1)	13/11	13.6 (12.8)
MET_2(11) Met31/Met32	AAAACTGTGG[b]	ACYSKG(4) CTGTGG(2)	39/8	4.8 (9.2)
PHO(5) Pho4	CACGTKNG[a]	ACGTGS(1) ACGTGC(1)	18/5	12.1 (7.2)
PDR(7) Pdr1/Pdr3	TCCGCGGA[b]	CCGYGG(1) CCGTGG(1)	18/4	15.3 (12.9)
INO_1(10) Ino2/Opi1	CATGTGAAWT[b]	CATGTG(1) CATGTG(1)	15/9	7.9 (10.5)
INO_2(10) unknown	–	CAACAA(2) CAACAA(2)	28/10	6.9 (5.2)
TUP(25) Mig1	$KANW_4ATSYG_4W$[b]	GYGGGG(1) GGGGTA(1)	33/18	11.7 (8.3)
YAP_1(16) Yap1	TTACTAA[b]	MTTASK(1) CATTAC(2)	99/16	5.1 (6.5)
YAP_2(16) unknown	–	CGTTCC(2) CGTTCC(1)	15/16	6.3 (8.0)
GAL(6) Gal4	$CGGN_5WN_5CCG$[b]	– –	–	–
HAP(8) Hap2/Hap3/Hap4	YCNNCCAATNANM[a]	– –	–	–
GCN(38) Gcn4	RTGACTCATNS[a]	TGACTC(1) TGACTC(1)	44/26	12.5 (7.1)
MAT(10) Matα2	CRTGTNNW[a]	CATGYA(3) CATGTA(2)	21/7	5.1 (6.3)

Table 2.1: Similarity of known and predicted binding sites in upstream regions of genes belonging to families of 11 yeast *Saccharomyces cerevisiae* transcription factors. Experimentally verified consensus sequences are taken from Transfac[a] [Wingender et al., 1996] or from van Helden et al. [1998][b].

Family name (no. of genes), bound factors	Experimentally verified motifs	Predictions: degenerated and normal alphabets (rank)	No. of motifs /seqs. with match	Z-score (Information content [bits])
c-Myc(4) c-Myc/Max	CACGTG[a]	– –	–	–
Ras(5) unknown	–	CGARCG(1) CGAGCG(1)	9/4	10.1 (10.9)

Table 2.2: Results of the frequency score calculation in two human families. The experimentally verified consensus sequence is taken from Transfac[a] [Wingender et al., 1996].

1998], RSA-tools-oligo-analysis [van Helden et al., 1998], and MEME version 2.2 [Bailey and Elkan, 1994] were also applied to detect motifs in the 11 yeast regulatory families. No algorithm was able to detect all of the described motifs.

In order to evaluate the applicability of the frequency score for the prediction of functional sites in the human genome, the algorithm was applied to human sequences (Tab. 2.2). No significant motif similar to the expected element CACGTG was found in the promoters of genes regulated via c-Myc. In contrast, the analysis of five genes regulated via H-Ras revealed the sig-

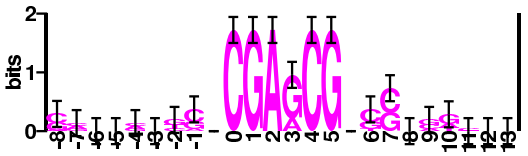


Figure 2.3: A sequence logo [Schneider and Stephens, 1990] of the significant pattern CGARCG found in promoters of human genes regulated via H-Ras. Bars correspond to sample size error corrections.

nificant pattern CGARCG (Z-score=10.1, information content=10.9 bits, see Fig. 2.3). The motif does not match any previously characterized site listed in the Transfac database [Wingender et al., 1996] and does not self-overlap in the promoter sequences matched.

2.3.2 Frequency score: variation of the parameters

In order to verify the motif predictions and to test the robustness of the algorithm, analyses were repeated with different sets of initial parameters. The calculations were performed with the nucleotide alphabet `ACGT`. In each run, a single parameter was changed while all other parameters were kept at their default values (motif length 6, upstream region of 800 bp, Markov model of order 3).

Family	Motif	4	5	6	7	8
NIT	GATAAG	4.5(1)	7.0(1)	**13.9**(1)	12.0(2)	10.0(15)
MET_1	TCACGTG	5.0(1)	6.1(1)	13.7(1)	23.7(1)	**24.3**(1)
MET_2	AAAACTGTGG	2.4(4)	4.4(3)	5.0(2)	10.6(2)	**15.6**(3)
PHO	CACGTKNG	5.8(1)	9.4(1)	10.3(1)	12.9(1)	**16.8**(1)
PDR	TCCGCGGA	6.2(1)	8.1(1)	11.8(1)	16.0(2)	**24.2**(1)
INO_1	CATGTGAAWT	1.7(12)	3.0(10)	7.9(1)	11.0(1)	**19.3**(1)
INO_2	aCAACAAs*	4.1(1)	5.2(1)	**6.9**(2)	6.8(7)	**6.9**(40)
TUP	KANWWWWATSYGGGGW	–	8.7(1)	10.2(1)	11.3(1)	**12.7**(2)
YAP_1	TTACTAA	1.8(8)	4.3(2)	5.1(2)	7.7(1)	**9.9**(2)
YAP_2	cCGTTCCs*	1.6(16)	3.3(4)	6.3(1)	6.5(5)	**9.1**(5)
GCN	RTGACTCATNS	4.5(1)	7.5(1)	12.5(1)	12.8(1)	**13.5**(1)
MAT	CRTGTNNW	2.0(8)	4.0(4)	5.1(2)	**6.8**(4)	6.0(38)

Table 2.3: Z-scores of predicted motifs as a function of the searched motif length (4, ..., 8). Only complete or shifted by 1 bp matches to previously identified consensus sequences were considered. Z-scores of the most over-represented oligonucleotides matching known consensus sequences are presented besides the ranks of the motifs (in brackets). *Patterns marked with a star represent new motif candidates.

Applying different word lengths revealed highly ranked motifs that match previously identified consensus sequences for most analyzed patterns (see Tab. 2.3 for results regarding the YAP and INO families). In this analysis, only complete matches to previously identified consensus sequences were considered (shifts of motifs by a maximum of 1 bp were allowed). When word lengths of 5 to 7 were selected, the top predictions for 6 regulatory families revealed complete matches to previously identified transcription factor binding sites. Choosing larger or smaller motif lengths led to worse predictions of corresponding known sites, but often provided additional information for motifs with wider conserved cores (e.g. the pattern `CATTACTAA` of the YAP family).

Moreover, the predictions were analyzed for different background models

Family	Motif	E	B	M1	M2	M3	M4	F
INO_1	CATGTG	5.4(17)	6.3(4)	6.6(2)	6.5(2)	**7.9**(1)	7.5(1)	7.7(1)
INO_2	CAACAA*	**12.0**(2)	9.7(1)	7.8(1)	8.4(1)	6.9(2)	5.9(2)	6.0(2)
YAP_1	CATTAC	5.9(25)	4.3(16)	5.4(2)	**6.5**(1)	4.5(3)	5.2(2)	5.0(2)
YAP_2	CGTTCC*	3.5(72)	6.6(5)	**6.8**(1)	6.3(2)	6.3(1)	6.5(1)	6.4(1)

Table 2.4: Z-scores (and their ranks) of motifs best matching the experimentally known consensus sequences depending on the background model (E = equiprobable base distribution; B = single nucleotide probabilities/Markov model order 0; M1, ..., M4 = Markov chain models of the orders 1 to 4; F = probability of motifs based on the 6-mer frequency in the training set/Markov model order 5). *Patterns marked with a star represent new motif candidates.

(results regarding the YAP and INO families are presented in Tab. 2.4). Assuming equal and independent probabilities of all nucleotides typically lower Z-scores were observed. When instead single nucleotide probabilities were considered, or when a Markov chain model of the order 1 was applied, the resulting predictions were better, but still biased towards non-specific, partly self-overlapping patterns. Applying Markov chain models of the orders 2 to 5 revealed the highest numbers of motifs that match previously identified consensus sequences.

Finally, the robustness of the predictions was tested for variations of the analyzed sequence length upstream of the translation start. It could be observed that there exist motifs, which have significantly higher scores for particular lengths of the upstream regions. These peaks of Z-score indicate strong positional preferences of particular motifs. For instance, the pattern CATGTA matching a known regulatory element of the MAT family revealed the highly significant Z-score of 7.7 and the top rank for an upstream region length of 400 bp, while a score of only 5.1 and the rank 2 resulted applying the default length of 800 bp.

Generally, the analysis of varying upstream sequence lengths indicates that most functional elements of yeast can be detected, when 800 bp upstream of the translation start are analyzed (see Tab. 2.5).

2.3.3 Combination of both scores

Scatter plots are a convenient representation of results of regulatory element search obtained with the ITB algorithm. For each studied motif the frequency score and the positional score constitute coordinates of a point on a

Family	Motif	400	500	600	700	800	900	1000
NIT	GATAAG	13.2(1)	**16.9**(1)	15.8(1)	15.1(1)	13.9(1)	13.5(1)	12.6(1)
MET_1	CACGTG	12.0(1)	13.5(1)	**14.7**(1)	**14.7**(1)	13.7(1)	12.7(1)	12.0(1)
MET_2	CTGTGG	**5.6**(3)	4.7(4)	5.5(3)	5.5(2)	5.0(2)	4.4(8)	4.0(9)
PHO	GCACGT	8.8(1)	10.0(1)	**12.1**(1)	11.0(1)	10.3(1)	10.3(1)	9.7(1)
PDR	CCGTGG	8.3(1)	8.6(1)	11.4(1)	11.6(1)	**11.8**(1)	11.0(1)	10.3(1)
INO_1	CATGTG	**10.5**(1)	9.2(1)	8.1(1)	7.3(1)	7.9(1)	8.5(1)	7.9(2)
INO_2	CAACAA*	2.9(20)	3.1(23)	3.6(11)	4.4(5)	6.9(2)	7.5(2)	**8.1**(1)
TUP	GGGGTA	5.2(3)	5.5(2)	**10.2**(2)	9.2(2)	**10.2**(1)	9.8(1)	9.1(1)
YAP_1	CATTAC	2.8(28)	4.4(6)	4.5(6)	**5.4**(2)	5.1(2)	4.4(3)	5.2(2)
YAP_2	CGTTCC*	**6.4**(1)	6.1(2)	5.3(4)	5.8(1)	6.3(1)	5.7(1)	6.2(1)
GCN	TGACTC	**16.8**(1)	15.9(1)	14.5(1)	13.4(1)	12.5(1)	11.4(1)	11.1(1)
MAT	CATGTA	**7.7**(1)	6.6(1)	5.8(2)	5.1(2)	5.1(2)	4.6(3)	4.2(4)

Table 2.5: Z-scores (and their ranks) of yeast motifs predictions as a function of the analyzed sequence lengths (in bp, upstream of the translation start site). *Patterns marked with a star represent new motif candidates.

scatter plot. Fig. 2.4 presents results of analysis for five yeast *Saccharomyces cerevisiae* gene families and a set of human genes regulated by the H-Ras. Motifs of length six constructed out of the standard nucleotide alphabet ACGT were enumerated in this analysis. Signals of motifs corresponding to known consensus sequences often appeared to have large scores and therefore to be located at the top right corner of the plots. Moreover, sequences similar to the expected patterns were frequently detected with high scores. When the MET, GCN, INO, TUP and NIT families were analyzed, signals corresponding to previously identified motifs represented the most appealing predictions clearly separated from the other spots. In case of the PDR and PHO families, several variations of the expected sequences TCCGCGGA and CACGTKNG were predicted – at least one of each having both significant Z-scores and positional scores.

Moreover, ITB extracted new candidates for transcription factor binding sites. For instance, an analysis of the PHO family revealed the significant candidate CGTATA (Z-score 4.4, positional score 3.2). However, significant positional scores are often caused by strongly self-overlapping signals. In the MAT family, the three highest-scoring motif candidates represent shifted variations of the strongly self-overlapping GACGAC. The previously identified motif CATGTA represents the most appealing prediction of this family, if the former three signals are not considered. The new candidate CAACAA of the INO family revealed lower scores than the expected CATGTG. However, a cal-

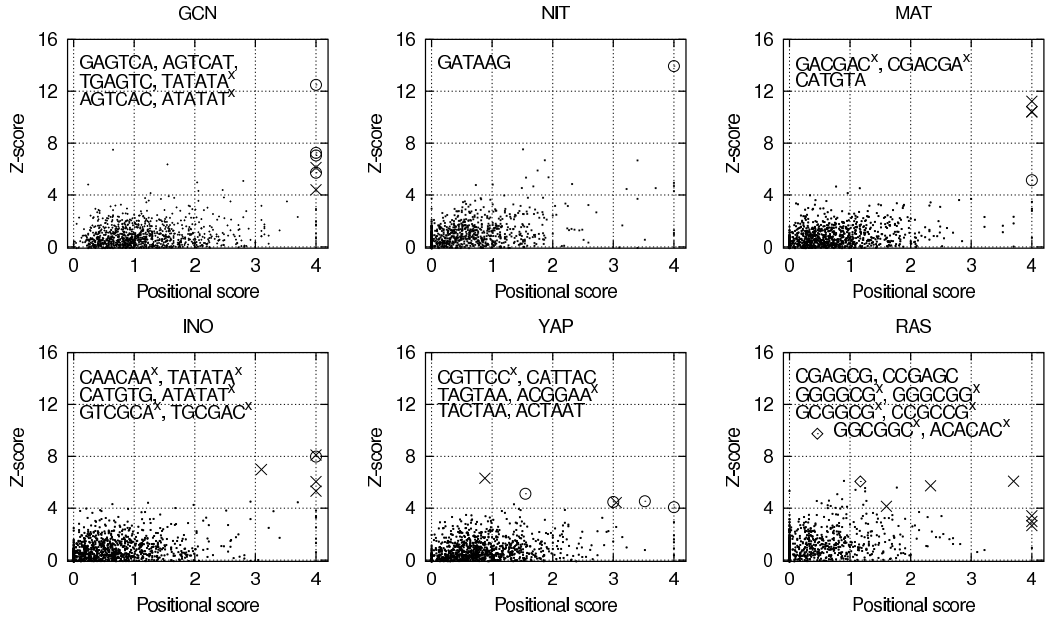


Figure 2.4: Scatter plots combining the frequency Z-score and the positional score calculated by ITB (using the nucleotide `ACGT` alphabet). Five yeast *Saccharomyces cerevisiae* co-regulated gene families and human genes co-regulated via H-Ras were analyzed. Circles correspond to motifs matching previously identified consensus sequences, while crosses ('x') indicate new candidates for sequences recognized by the transcription factors. Diamonds indicate oligonucleotides matching the degenerated motif `CGARCG`, which had been found independently by the frequency score calculation procedure when the alphabet containing degenerated bases was used. Oligonucleotides are listed from the left to the right (or – if signals are one upon the other – from the top to the bottom). Self-overlapping motifs were not removed from the output.

culated Z-score of 6.9 along with a significant positional score of 3.1 stress the potential importance of the candidate. In the case of the YAP family motifs representing good matches to the expected motif `TTACTAA` or the potential candidate `CGTTCC` were found at the top right corner.

Analyzing the human genes co-regulated via H-Ras revealed a number of potential motif candidates. However, most of those signals represent strongly self-overlapping sequences or rather non-specific sequences like `GGGCGG` or `GGGGCG`. The sequence `CCGAGC`, a good (shifted) match to the previously identified candidate `CGARCG` (see Fig. 2.3), revealed a significant Z-score of 6.1 along with a positional score of 1.2.

2.4 Discussion

ITB (Integrated Tool for Box finding introduced in section 2.2) is a sensitive and powerful algorithm integrating frequency and positional information to detect transcription factor binding sites in promoters of co-regulated genes. The algorithm performs an exhaustive search for regular expression-like patterns, allowing matching of more than one nucleotide at any position a motif. Such an approach may be considered as a good compromise between searches for frequent oligonucleotides and weight-matrix based methods. It allows the detection of motifs that are not completely conserved – and it guarantees to find the most significant elements due to the exhaustive search strategy. In order to correct an enlarged variance of the number of motifs due to self-overlapping, an appropriate correction formula is applied (Eq. 2.2). Sinha and Tompa [2000] provide an analysis of the algorithm complexity as the function of the alphabet size and the length of studied motifs.

Based on the performed parameter variations motifs of lengths six or seven nucleotides were found as a reasonable choice for performing searches with ITB in the yeast *Saccharomyces cerevisiae* genome, although analyzing wider sequences may still provide further information. Modeling background probabilities with Markov models of orders from two to five led to comparable predictions. Variation of upstream sequence lengths led to significant changes of the frequency scores, indicating positional peculiarities of yeast motifs. ITB detected highly significant motifs corresponding to functional regulatory elements found by experimental analysis. Only a limited set of additional patterns was predicted. Even when positional information was not considered, known regulatory elements of 8 out of 11 yeast regulatory families were predicted correctly using the alphabet allowing matches of multiple nucleotides. Similar results were obtained using the standard `ACGT` alphabet.

Combining both the frequency score (section 2.2.1) and the positional score (section 2.2.2) increases the specificity of ITB. Motifs with significant positional scores, which correspond to previously identified consensus sequences, were extracted from 9 out of 11 families (see section 2.3.3). Only few new candidate motifs having both high positional and frequency scores were predicted. A couple of reasons might lead to the observed positional preferences. Among these are interactions of transcription factors with the pre-initiation complex, the removal of single nucleosomes within a promoter, protein-protein interactions within factors that stabilize weak protein-DNA interactions, or recent duplications of regulatory regions. Moreover, an accumulation of functional sites might serve to increase the local concentration of transcription factors.

ITB failed to detect described consensus sequences of two yeast fami-

lies (presented in Tab. 2.1). Some tools like AlignACE, Dyad-detector [van Helden et al., 2000b] or the method in Sinha and Tompa [2000] are capable of predicting the consensus described for the GAL family. ITB does not detect sequences like the Gal4 binding site containing longer gaps.

Some motifs predicted by ITB are potential candidates for novel yeast transcription factor binding sites. The elements `CGTTCC` and `CAACAA` represent strong candidates for regulatory elements involved in the regulation of the YAP and INO families. Moreover, analysis of the PHO family revealed a significant candidate `CGTATA`. These predictions are highly robust against variations of the algorithm parameters. Furthermore, the calculation of high positional scores for the candidates or at least for other motifs matching the candidates supports a selection of these patterns. Searching Transfac [Wingender et al., 1996] for the elements `CGTTCC`, `CAACAA` and `CGTATA` did not reveal matches to known yeast consensus sequences.

Since ITB succeeded in detecting regulatory elements in yeast upstream regions, it might sound reasonable to assume that this method may also work in human promoters. This chapter presents results for two human gene sets. The expected binding site was not detected for genes controlled by the c-Myc transcription factor, but the motif `CGARCG` appears as a candidate transcription factor binding site of genes co-regulated via H-Ras. However, the mechanisms regulating transcription in higher eukaryotes are significantly more complex. The ITB algorithm assumes a direct interaction of a single transcription factor with upstream regions close to the genes. However, in higher organisms a cooperation of many factors needs to be taken into account (more details in chapter 5). Moreover, active promoters are often located far from protein coding sequences. Unfortunately, only a limited number of experimentally verified promoter sequences is available [Perier et al., 1998, Suzuki et al., 2002], therefore prediction algorithms might have to be used. Additionally, larger gene regulatory networks make direct and indirect targets more difficult to distinguish. Usage of proper training sets and the reliable detection of co-regulated genes are prerequisites for an analysis of regulation in the genomes of higher eukaryotes [Herzel et al., 2001]. Since these problems cannot be left aside, applications of several algorithms (including ITB) applied to human sequences are studied in detail in chapter 3.

Chapter 3

Regulatory elements of AP-1 and RAS targets

Summary

This chapter contains a description of a computational tool designed for prediction of cis-regulatory elements. The method combines predictions of different algorithms: Clover, Cluster-Buster, an own implementation of human/rat/mouse sequence identity and the ITB algorithm (described in chapter 2). The procedure utilizes data from the human genome sequence, NCBI gene annotations, the database of verified eukaryotic promoters (EPD), the collection of experimentally proven binding sites (Transfac) and homologies of human, mouse and rat (HomGL/HomoloGene).

Following Kielbasa et al. [2004b], two applications of the tool are discussed. First, as presented in section 3.2, the method is tested on a collection of 18 upstream regions of experimentally verified AP-1 target genes. About a half of the known sites belongs to the high-scoring candidates found by the tool.

Next, the same analysis is applied to genes found to be up- or down-regulated as an effect of RAS transformation in rat fibroblasts (section 3.3). A detailed literature and computational search for promoter regions of these genes has been performed. Indications of overrepresented matches to the motifs recognized by Elk-1 and AP-1 transcription factors are found via a comparison with shuffled promoter sequences. In some promoters consistent predictions of clustered binding sites were obtained.

3.1 Introduction

Growing databases of mRNA expression profiles increase the need for bio-computational methods to predict regulatory rules in sets of co-regulated genes. A typical analysis of a large scale microarray experiment results in clusters of genes sharing similar expression profiles as functions of different experimental conditions or time points [Herzel et al., 2001]. Emergence of such patterns might be a hint suggesting existence of regulatory mechanisms controlling genes within each of the clusters.

Yeast *Saccharomyces cerevisiae* was one of the first organisms, whose genome was studied using the approach mentioned above. Several groups [van Helden et al., 1998, Pilpel et al., 2001, Caselle et al., 2002] predicted or verified short DNA sequences bound by transcription factors in the upstream regions of genes assigned to a common cluster (see Tab. 2.1). Problems arise when these methods are directly applied to higher eukaryotic organisms. The number of incorrectly predicted binding sites grows with the total length of DNA sequences in which the search is performed. This observation emphasizes the role of careful preselection of short promoter regions (instead of full upstream sequences in yeast) for further search.

Furthermore, one cannot ignore local properties of the DNA sequences selected. Noticeable changes of GC content (CpG islands versus CpG suppression) or poly–A sequences, known to occur frequently close to transcription start sites, may dominate the scoring procedure and produce unexpected results.

Certainly, to increase prediction quality all potential sources of information should be combined. Here a method is presented, which in the analysed set of co-regulated promoters associates sites matching known profiles recognized by transcription factors with statistically overrepresented motifs. Additionally, since it is probable that the regulatory mechanisms are preserved in the course of evolution [Wasserman et al., 2000, Dieterich et al., 2002], the results are presented in the context of promoter homologues between human, rat and mouse.

Last but not least, the presented schema of the regulatory motifs detection requires compilation of information stored in several databases: Ensembl for human, mouse and rat [Birney et al., 2004], RefSeq [Pruitt et al., 2005], EPD [Perier et al., 1998], DBTSS [Suzuki et al., 2002], Homologene, Unigene [Wheeler et al., 2003], Transfac [Wingender et al., 2000, Matys et al., 2003]. Even though the task seems to be well defined, the practical implementation is difficult due to different standards for expressing gene names and base positions within genomic sequences. In order to overcome these technical issues, a dedicated database HomGL has been developed by Blüthgen et al.

[2004], which provides a direct access to homologues and their upstream regions, as well as it handles a variety of gene accession identifiers.

3.2 Study of AP-1 regulated genes

In order to elucidate properties of the prediction algorithms an artificial collection of genes was constructed. The table `genes` of the Transfac database was scanned for genes regulated by the AP-1 transcription factor. Only human genes were selected, based on the `HS$` pattern present in the gene identifier. Next, these genes were identified using the human Ensembl service and the corresponding DNA sequences were extracted. Additionally, with the help of the HomGL database the human genes were mapped and their mouse and rat homologues were located. Fragments from 1500 bp upstream to 200 bp downstream around transcription start site (as provided by Ensembl) were used.

The table `sites` of the Transfac database was processed to find the short DNA sequences representing the binding sites experimentally proven to be recognized by the AP-1 transcription factor. In 18 cases it was possible to match exactly the substring representing the factor binding site within the extracted human DNA sequences. It should be noted, that there were no data available stating whether an extracted site was a unique AP-1 site per gene or the one with the highest affinity.

The final piece of information taken from the Transfac database came from the `matrix` table. Since more than one positional frequency matrix was provided for the AP-1 transcription factor, the one was chosen which corresponded to the largest number of sequences aligned (a matrix identifier `M00174`, an alignment of 56 observations). This single matrix was used for prediction of binding sites.

Using this data set, results of the following algorithms were studied:

- the Clover program [Frith et al., 2004a] calculating a similarity scores between a weight matrix and DNA sequences belonging to a family (several advanced background models might be taken into account);

- the Cluster-Buster program [Frith et al., 2003] detecting multiple close occurrences (clusters) of binding sites predicted based on weight matrices from a provided set;

- an own implementation of a percent identity [Schwartz et al., 2000] procedure reporting percent of similarity between homologous DNA sequences;

Gene	Clover	Cluster-Buster	%-identity
Hs.511899	1	+	h
Hs.1832	1	+	h
Hs.435800	1	+	m
Hs.25647	1	−	0
Hs.694	1	−	0
Hs.89679	3	−	0
Hs.418083	4	0	0
Hs.1349	4	−	0
Hs.202453	5	−	l
Hs.253495	6	0	m
Hs.856	6	−	0
Hs.408312	7	−	h
Hs.78465	7	−	m
Hs.385521	10	−	l
Hs.446641	> 10	0	0
Hs.1905	> 10	−	m
Hs.83169	> 10	−	h
Hs.375129	−	0	h/m

Table 3.1: Predictions of AP-1 binding in gene upstream sequences experimentally proven to contain AP-1 binding sites. Clover: rank at which the known site was predicted ("−" stands for no prediction at all). Cluster-Buster: "+" the known site within a predicted cluster; "−" the known site outside a predicted cluster; "0" no cluster was predicted. %-identity: "h" – high identity at the experimental site, "m" – average, "l" – low, "0" – no data available.

- the ITB algorithm (discussed in chapter 2, Kielbasa et al. [2001]) calculating overrepresentation of short DNA words (5,6 or 7 nucleotides) in a family of DNA sequences in comparison to a training (background) set of DNA sequences.

The Clover program was requested to perform calculations with all three background methods it provides (mononucleotide DNA sequence shuffling, dinucleotide shuffling, and shuffling of the positions of the weight matrix). The score thresholds were lowered (min. motif score 3.0, max. P-value 0.15), since with the default values majority of the experimentally known sites were not found. A similar approach was used with the Cluster-Buster program – the motif score threshold was reduced to 3.0 and the cluster score threshold to 2.5. The ITB algorithm was run with the shuffled AP-1 sequences as the

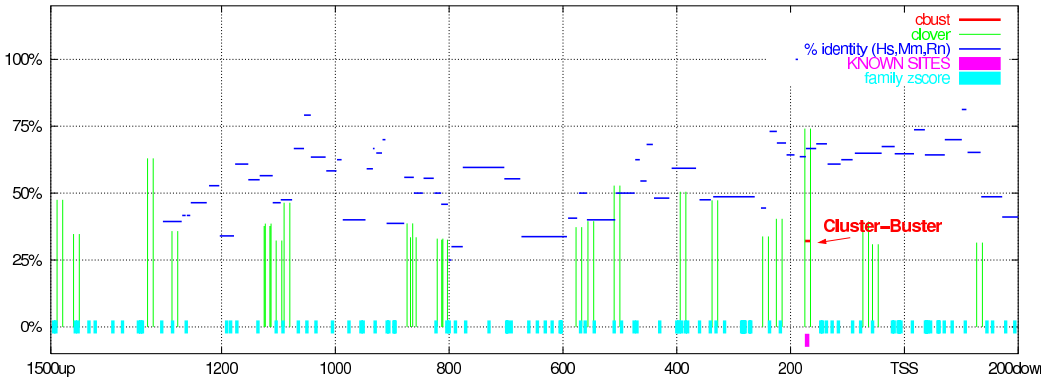


Figure 3.1: Predictions of AP-1 binding sites in upstream sequence of gene Hs.511899 (EDN1, endothelin 1). The single bar at the bottom of the picture shows the position of an experimentally proven AP-1 binding site. Clover: predictions are shown with thin vertical lines (height is proportional to similarity score). Cluster-Buster: the single prediction is marked by the arrow. The broken horizontal lines represent ungapped stretches of the sequence similar to rat and mouse (similarity scale provided on the vertical axis). ITB: the bars on the horizontal axis mark locations of the mostly overrepresented 6-mers.

background set and for motif lengths 5, 6 and 7. No significant overrepresentation of DNA words was detected by the ITB algorithm. A detailed analysis of the AP-1 sites revealed that there was no significant excess of 5-mers due to the considerable variety of the sites.

The results of the mentioned programs as well as experimentally known sites were summarized together on graphs like in Fig. 3.1, describing each DNA sequence of the AP-1 collection separately. The summary for all 18 sequences is given in the Tab. 3.1. Even in this artificial set with enriched AP-1 sites and with a single weight matrix, only a half of the sites were detected by Clover (assuming a practical limit of top five candidate sites). The top-ranked Clover motifs belong to groups reported by the Cluster-Buster program, i.e. such clusters may be used as an additional indicator of true sites. The human, mouse, rat percent identity signals tend to have values higher than average at the experimentally known sites, and can be used also as an additional indication of true sites.

3.3 Study of the RAS-dependent genes

In order to find candidate genes involved in tumor development two cell lines have been studied by Zuber et al. [2000]: preneoplastic rat 208F fibroblasts and its malignant HRAS-transformed derivative FE-8. Subtractive suppression hybridization [Diatchenko et al., 1996] technique was used to find gene fragments up-regulated or down-regulated upon neoplastic transformation. Out of more than 1200 subtracted cDNA clones 244 have been recognized by the authors as already known genes.

For the purposes of this chapter, first human homologues of these genes were identified. Based on the human Unigene set and alignments performed with the Blast algorithm [Altschul et al., 1997] 216 human homologues were found. Afterward, based on the human genome assembly provided by Ensembl, upstream regions of the homologues were extracted using the HomGL system. DNA fragments from 1500 bp upstream to 200 bp downstream around the transcription start site (as reported by Ensembl) were selected for further analysis.

Several available promoter prediction programs: First Exon Finder [Davuluri et al., 2001], McPromoter [Ohler et al., 2001] and CONPRO [Liu and States, 2002] were applied to these sequences. Additionally, literature and database (Eukaryotic Promoter Database by Perier et al. [1998], NCBI databases) search for experimentally identified promoters of the genes was performed. After these steps the analysis was limited to the genes for which at least two of the programs predicted transcription start sites (TSS) not further than 300 bp from each other. Finally, promoters of 30 RAS suppressed genes and 20 RAS activated genes were selected.

The Transfac database (version 6.0, Wingender et al. [2000]) provided a collection of more than 300 (partially redundant) profiles recognized by vertebrate transcription factors. In contrast to the study of the genes regulated by the AP-1 transcription factor where only a single profile was studied (section 3.2), here all the profiles provided by the library should be separately analysed. Unfortunately, when searching for binding sites the amount of false predictions grows proportionally to the number of the profiles studied. Therefore, only the profiles mentioned in the literature in the context of RAS regulation were chosen for further analysis. In total 52 weight matrices were selected. Naturally, such a selection of profiles is biased towards the knowledge of the user performing the literature search. An unbiased method which compares the statistical properties of the profiles in order to select nonredundant subset of them is discussed in chapter 4.

The same prediction algorithms as used in the AP-1 study were applied to the RAS suppressed and the RAS activated sets of promoter sequences

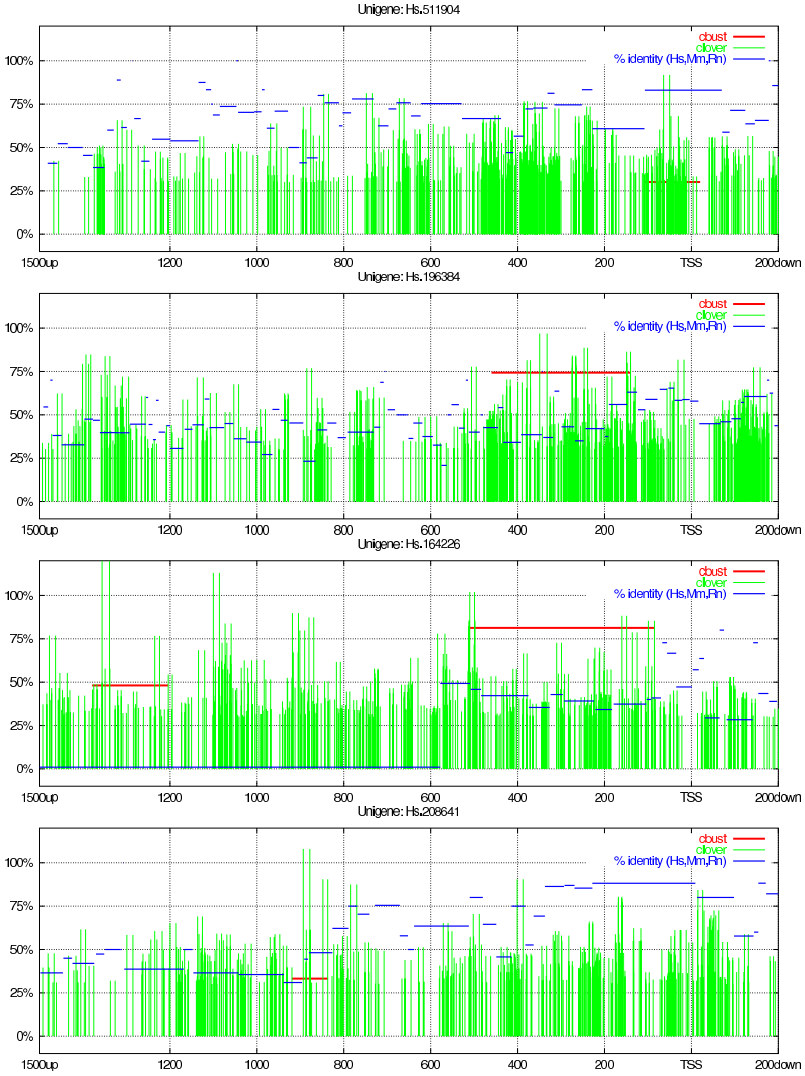


Figure 3.2: Examples of predictions of binding sites in promoters of genes classified as suppressed by RAS. The two top charts (genes PTGS2 and EIF4A2) present regions of higher human-mouse-rat homology and clustered occurrences of predicted binding sites: E2F, TBP (in PTGS2); NF-κB, AP-4, Sp1, E2F (in EIF4A2). The third graph (THBS1 gene) shows a region with a cluster, but with a lower homology level. On the bottom graph (ACTA2 gene) the regions of high homology and good clustering are mutually exclusive.

for all the selected profiles (Fig. 3.2 demonstrates results for four promoters). The number of Clover hits increased significantly in comparison to the AP-1 study due to the larger number of studied profiles. Nevertheless, in the whole set of 50 promoter sequences only 6 regions with strong signals given by all three algorithms were found. In the case of AP-1 study 3 out of 8 strong signals given by all methods were related to the experimentally verified sites. Therefore it is likely that similar fraction of the six cases of RAS target genes represent true signals.

Overrepresentation of binding sites in promoters

An additional method studied for the sets of the RAS target genes considered the possibility that the whole set of genes is a target of the same transcription factor. In such a case the distribution of weight matrix match scores should be different in the original sequence set compared to a random sequence set. The relative entropy score was used to model factor affinity to a position of a DNA sequence [Schneider et al., 1986]. The background probabilities were counted locally in a sliding window of 201 bp centered at the studied position.

Predicted transcription factor binding sites depend strongly on the chosen minimum score threshold. Here, in order to study the general case the number of predicted binding sites is counted as a function of varying thresholds. In Fig. 3.3 the average distance between predicted binding sites is plotted as a function of the score threshold. Consequently, small distances correspond to many detected binding sites. The thick lines present the results obtained for binding sites predicted in the original RAS promoter sets. Additionally, results of the same analysis applied to several shuffled promoter sequences are drawn as thin lines. Two methods of shuffling have been studied: "horizontal" – reordering nucleotides within each promoter separately and "vertical" – exchanging the nucleotides at same positions in TSS-aligned promoters (so this method preserves the GC content as the function of the distance to TSS). Thus, the thick line appearing below the thin lines indicates overrepresentation of binding sites corresponding to the studied transcription factor. This information might be used for preselection of matrices chosen from a large library for more detailed binding studies.

3.4 Discussion

The study presented in the previous sections describes a tool for prediction of binding sites in promoters of human genes assumed to be regulated by the

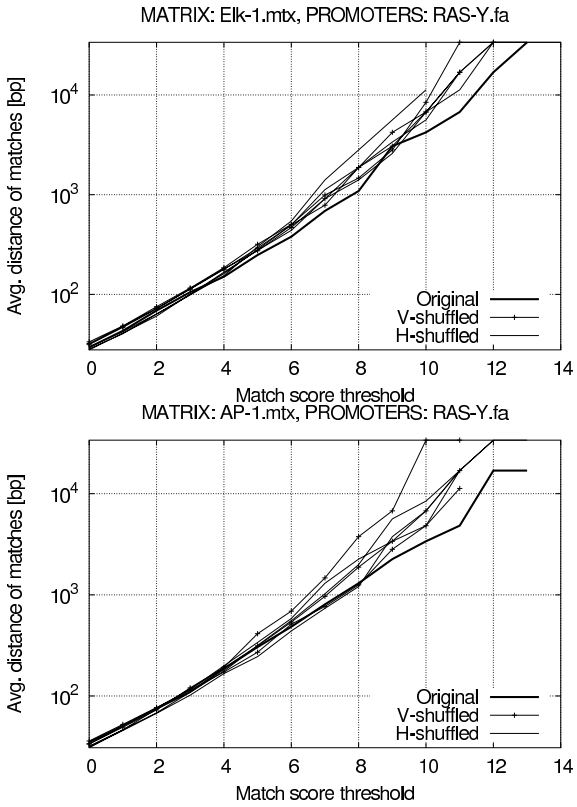


Figure 3.3: The average distance between predicted binding sites of transcription factors: Elk–1 (above), AP–1 (below) as a function of the minimum score threshold. The thick lines show results for the original promoter sequences, the thin lines for these sequences randomly shuffled.

same transcription factor. The HomGL system is used to extract upstream sequences of the genes as well as the upstream regions of their mouse and rat homologues. Afterward regulatory element prediction programs are executed and their results are presented on charts for each processed gene.

In order to evaluate the prediction quality the method was studied using an artificial data set of AP-1 target genes (section 3.2). Based on annotations provided by the Transfac database a list of 18 genes with known sites bound by the AP-1 transcription factor was compiled. The analysis performed with the binding sites prediction algorithms resulted in eight regions in the upstream sequences where high scores were reported by all the programs. Three of the regions corresponded to the experimentally known sites. The other five might represent false positives or other binding sites not listed in the database. The same strategy was applied to two sets of RAS target genes

(section 3.3). Here, the algorithms indicate six candidate regions in 50 genes, which might contain clusters of regulatory elements.

The studied examples illustrate that different binding site prediction techniques can be combined to improve the reliability of transcription factor binding site predictions. In the current implementation of the method three features are combined. Profiles recognized by transcription factors are used to search for highly scoring and overrepresented candidate binding sites in the gene upstream sequences. Besides, clustering of the predicted sites is analysed as well as their location in the evolutionary conserved regions of the sequences.

The study suggested several issues where improvements are necessary in order to achieve better quality of the predictions. First, it is highly probable that the extracted short upstream sequences do not contain the corresponding promoters, and therefore the searched regulatory elements are not present in the input sequences. Therefore, a reliable source of regions containing promoters is needed.

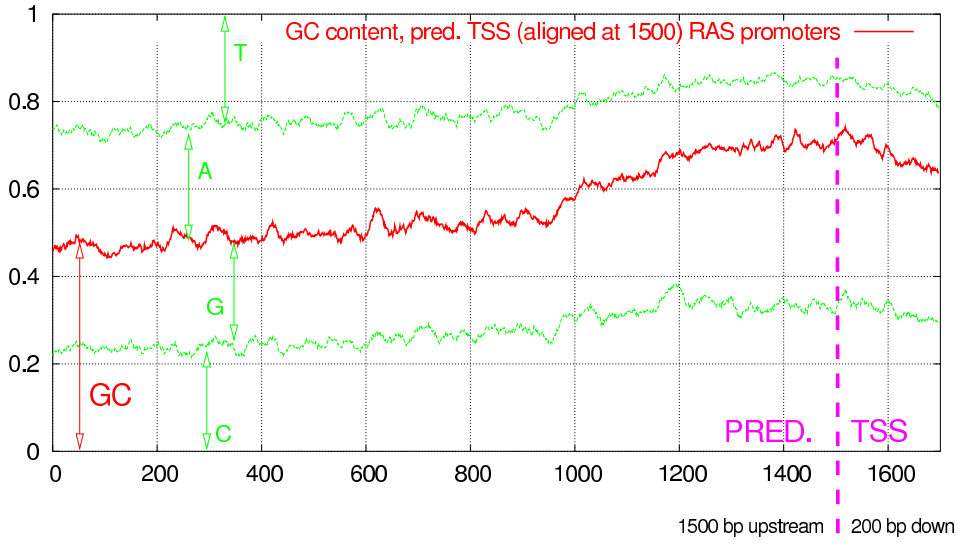


Figure 3.4: GC-content of the promoters the RAS-dependent genes. The promoters have been aligned at the predicted (or known, where available) transcription start site. For each position from 1500 upstream to 200 downstream the fractions of each nucleotide have been calculated and averaged over the surrounding 5 positions downstream and 5 upstream.

Second, the upstream regulatory regions are large and quite heteroge-

neous with respect to their composition. Fig. 3.4 demonstrates the nucleotide GC-content in the studied set of RAS promoters. They have been aligned at the predicted transcription start sites. The picture shows the fractions of each of the nucleotides calculated over all promoters as a function of the distance to the predicted transcription start sites. Close to the TSSes, the proportion of `C` and `G` nucleotides grows approximately to 70% in comparison to less than 50% at distances larger than 1000 bp. When this observation is ignored, the predicted binding sites rather reflect the GC-content of the profiles recognized by transcription factors, instead of the specific profile patterns. Consequently, in calculation of the binding scores local background models should be preferred in order to minimize artifacts due to heterogeneous composition of promoters.

Next, although some of the known AP-1 binding sites share all three studied properties, it is clear that majority of the sites remained either undetected or dominated by false signals. Definitely the signals studied here are not predictive enough to provide results of such quality which is needed in design of experiments. Only a careful combination of diverse informations can lead to a successful prediction. Since transcription factors rely on subtle signals of different nature, their detection require an approach combining multiple sources in a probabilistic way [Pudimat et al., 2004].

Therefore in the following two extensions are discussed. In chapter 4 a method for selection of independent transcription factor binding profiles is presented. This approach allows to group similar profiles and choose only representatives in each of the groups. This way a preselection of profiles, which is typically necessary for further analysis, can be performed on a statistical basis, which reduce a potential bias introduced by a manual selection. Additionally, a novel method is presented in chapter 5. This technique uses growing functional gene annotations and gives a possibility to predict biological functions of combinatorial interactions of transcription factors.

Chapter 4

Similarities of profiles recognized by transcription factors

Summary

Redundancies present in the libraries of profiles recognized by transcription factors decrease the quality of binding site predictions. Low average information content of the profiles and their high number limit the practical usage of simple profile scoring methods, due to a large rate of false positive signals. Therefore, as discussed in chapter 3, additional information need to be taken into account. As presented in section 3.3 a careful selection of independent transcription factor profiles is needed, but a bias caused by a manual profile selection should be eliminated. This chapter proposes a method for identification of sets of similar profiles, as discussed in Kielbasa et al. [2005]. The profiles which predict nearly the same binding sites are grouped into clusters, on the basis of two independent similarity measures. The first one compares nucleotide frequencies along all positions of two profiles, using the χ^2 distance measure (section 4.2.2). The second measure (section 4.2.3) correlates sets of binding sites predicted for the compared profiles. Since typically the profiles are constructed out of small numbers of aligned sequences, both measures take these sample sizes into account. Properties of both measures are studied in section 4.3.1, and afterward they are applied to identify redundancies of Jaspar and Transfac profile databases (section 4.3.2).

4.1 Introduction

In order to dissect the complex machinery of transcriptional control computational tools are widely used [Wasserman and Sandelin, 2004]. Candidate binding sites of known transcription factors are located by consensus sequence search or binding scores calculated from position weight matrices (PWMs) [Stormo, 2000], based on the statistical mechanics theory analysis showing that the logarithms of the base frequencies should be proportional to the binding energy contributions of the bases [Berg and von Hippel, 1987]. These matrices are derived from position frequency matrices (PFMs) obtained by aligning nucleotide sequences of binding sites recognized by a given transcription factor. PFMs contain the observed nucleotide frequencies at each position of the alignment. A popular collection of eukaryotic PFMs is given by the Transfac database [Wingender et al., 1996, 2000, Matys et al., 2003]. Furthermore, an open-access database, Jaspar [Sandelin et al., 2004a], has been compiled recently.

Several on-line tools are available to calculate high-scoring sites on the basis of such matrix collections [Quandt et al., 1995, Kel et al., 2003, Frith et al., 2004a]. These models predict for a given transcription factor many binding sites in distances of about 1000 bp by chance implying a high excess of false positives [Wasserman and Sandelin, 2004]. The situation is even worse if hundreds of different binding profiles are studied in parallel leading to multiple testing issues. Then high-scoring predictions are obtained every few base pairs. Often these predictions overlap as a result of similarities of transcription factor binding profiles.

First steps to overcome the flood of false positive signals are accurate predictions of promoter regions and enhancers [Werner et al., 2003, Ohler et al., 2001, Davuluri et al., 2001], phylogenetic footprinting [Wasserman et al., 2000, Dieterich et al., 2002, Ureta-Vidal et al., 2003] or the incorporation of expression profiling [Bussemaker et al., 2001, Hughes et al., 2000]. Another helpful strategy is the *a priori* reduction of the number of matrices to be considered. However, a user-defined preselection of a few matrices is highly subjective and might hide novel interactions of several transcription factors. Therefore, in this chapter two objective criteria to measure similarities of transcription factor binding site profiles are combined and studied. These measures allow to group similar profiles and make redundancies in collections of binding site profiles visible.

Redundancies in the collections of matrices may arise from several sources:

1. Identical transcription factors are represented by different matrices. This appears, e.g., due to the distinct nomenclature in Jaspar and

Transfac (for example the TATA-binding protein is referred as TATA in Transfac and as TBP in Jaspar) or due to the availability of matrices obtained with different methods (see for example Transfac matrices `V$SRF_01` and `V$SRF_Q6`) or stringency criteria (see for example `V$AP1_Q2` and `V$AP1_Q6`).

2. Factors within one family are represented by similar matrices due to the conserved structure of DNA-binding domains [Sandelin and Wasserman, 2004]. For example, both ATF and CREB matrices belong to the same bZIP family and recognize the `TGACGT` consensus sequence.

3. There might be so far undetected similarities of different transcription factor binding sites. Such similarities can point to a possible cross-talk between different regulatory pathways (discussed in section 4.3.3).

In this chapter two similarity measures are combined in order to identify similar matrices. The first one, presented in section 4.2.2, is based on the χ^2 distance of nucleotide frequencies provided by position frequency matrices. The other measure uses the corresponding position weight matrices and it constructs vectors of match scores of the matrices along a test DNA sequence. The Pearson correlation coefficient of the score vectors is taken as the similarity measure of the weight matrices, as discussed in details in section 4.2.3. Although each of these similarity measures has been already studied individually [Haverty et al., 2004, Hannenhalli and Levy, 2002, Hughes et al., 2000, Sandelin and Wasserman, 2004, Pietrokovski, 1996], their combination (section 4.3.1) reveals that they capture different properties of the matrices and therefore they complement each other. Moreover, since for many matrices only few experimentally verified binding sites are available these small sample sizes are taken into account in the both measures. The final part of this chapter presents two applications: in section 4.3.2 matrices provided by Jaspar and Transfac databases are compared and a list of redundancies is presented (Tables B.1 and B.2). The second application cites mapping of CLOCK-BMAL1 binding sites of circadian clock genes to the Myc-Max family (section 4.3.3).

4.2 Methods

4.2.1 Jaspar and Transfac databases

This chapter discusses similarities and redundancies present in databases of profiles recognized by transcription factors. Transfac is a commonly used

(commercial) database of experimentally verified transcription factor binding sites and profiles constructed based on them [Heinemeyer et al., 1998, Wingender et al., 2000, Matys et al., 2003]. The release from May 2004 provides 694 position frequency matrices (PFMs) covering vertebrates, plants, insects and fungi. Moreover, a publicly available Jaspar database was compiled recently, containing 108 PFMs associated mainly to vertebrates [Sandelin et al., 2004a].

Property	Transfac	Jaspar
Number of original matrices	694	108
Number of matrices after filtering	637	103
Min length	6	6
Max length	30	30
Median length	12	11
Min sample size	5	6
Max sample size	389	389
Median sample size	18	23
Min information content	3.6	5.7
Max information content	44.3	26.2
Median information content	12.8	11.6

Table 4.1: Summary of matrices provided by Jaspar and Transfac databases. Matrices for which the sample size was normalized to 100 and no information about the actual number of samples was available, as well as matrices of length below 6 or sample size below 5 were removed.

There exists no unique notation for storing the profiles and the information associated with them vary. Therefore, for the purposes of the study described in the following sections, all matrices with inconsistencies (e.g. no specification of the number of sites aligned to construct a matrix) were discarded. Furthermore, rather poor matrices of lengths below 6 positions or sample sizes below 5 sites were excluded. After these consistency checks and filtering steps the Jaspar set contained 103 matrices, and the Transfac library 637 matrices (see Tab. 4.1). All the matrices can be characterized by their length, sample size, and information content [Schneider et al., 1986].

4.2.2 χ^2-based distance D between position frequency matrices

Counts of nucleotide occurences at subsequent positions along a binding site profile are the direct outcome of DNA sequences alignment algorithms. Assuming the independence of positions within an alignment, these counts constitute a position frequency matrix (PFM) for the profile. The similarity measure D expresses the number of statistically significant different positions between two given profiles, minimized over possible shifts of one profile over the other. In other words D tells how many positions, at least, significantly differ between two binding site profiles constructed out of two alignments.

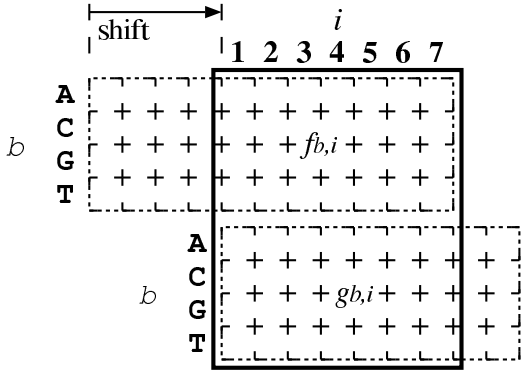


Figure 4.1: Variables used in the definition of χ^2-based distance measure D. Overlapping parts of two compared matrices f and g are compared. The calculation is performed for all shifts.

In the following $f_{b,i}$ and $g_{b,i}$ denote the entries of the overlapping parts of the two position frequency matrices under study, where b enumerates one of the four nucleotides `A`, `C`, `G` or `T` and i describes the position within the overlap (see Fig. 4.1). In order to test whether the counts of nucleotides at a position i are significantly different the homogeneity test using the χ^2 measure with 3 degrees of freedom might be used. The χ^2 distance at the position i is then given by

$$\chi^2 = \sum_{b=\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}} \frac{(N_{g,i} f_{b,i} - N_{f,i} g_{b,i})^2}{N_{f,i} N_{g,i} (f_{b,i} + g_{b,i})}, \tag{4.1}$$

where $N_{f,i} = \sum_b f_{b,i}$ and $N_{g,i} = \sum_b g_{b,i}$ are the sample sizes of the matrices columns at position i. In case of $f_{b,i} = g_{b,i} = 0$ for a certain b, the corresponding term is skipped in the sum. If χ^2 exceeds the threshold of

$\chi^2_{\text{th}}(p = 0.05) = 7.81$ the columns can be regarded as statistically different with a p-value of 0.05. Tab. 4.2 shows when the threshold is crossed for several typical nucleotide distributions f_b.

	$\mathtt{N}_{\text{weak}}$	$\mathtt{N}_{\text{strong}}$	$\mathtt{A}_{\text{weak}}$	$\mathtt{A}_{\text{strong}}$	$\mathtt{T}_{\text{weak}}$	$\mathtt{T}_{\text{strong}}$	$\mathtt{W}_{\text{weak}}$	$\mathtt{W}_{\text{strong}}$	$f_{\mathtt{A}}$	$f_{\mathtt{C}}$	$f_{\mathtt{G}}$	$f_{\mathtt{T}}$
$\mathtt{N}_{\text{weak}}$	$\cdot$	$\cdot$	$\cdot$	$\neq$	$\neq$	$\neq$	$\cdot$	$\neq$	2	1	2	1
$\mathtt{N}_{\text{strong}}$	$\cdot$	$\cdot$	$\neq$	$\neq$	$\neq$	$\neq$	$\cdot$	$\neq$	6	6	6	6
$\mathtt{A}_{\text{weak}}$	$\cdot$	$\neq$	$\cdot$	$\cdot$	$\neq$	$\neq$	$\cdot$	$\cdot$	6	0	0	0
$\mathtt{A}_{\text{strong}}$	$\neq$	$\neq$	$\cdot$	$\cdot$	$\neq$	$\neq$	$\neq$	$\neq$	24	0	0	0
$\mathtt{T}_{\text{weak}}$	$\neq$	$\neq$	$\neq$	$\neq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	0	0	0	6
$\mathtt{T}_{\text{strong}}$	$\neq$	$\neq$	$\neq$	$\neq$	$\cdot$	$\cdot$	$\neq$	$\neq$	0	0	0	24
$\mathtt{W}_{\text{weak}}$	$\cdot$	$\cdot$	$\cdot$	$\neq$	$\cdot$	$\neq$	$\cdot$	$\cdot$	3	0	0	3
$\mathtt{W}_{\text{strong}}$	$\neq$	$\neq$	$\cdot$	$\neq$	$\cdot$	$\neq$	$\cdot$	$\cdot$	12	0	0	12

Table 4.2: The symbol "$\neq$" marks pairs of nucleotide distributions reported by the χ^2 measure as significantly different. The "$\cdot$" symbol marks pairs, which do not cross the $\chi^2_{\text{th}}(p = 0.05) = 7.81$ threshold. The f_i values list nucleotide counts used in the test.

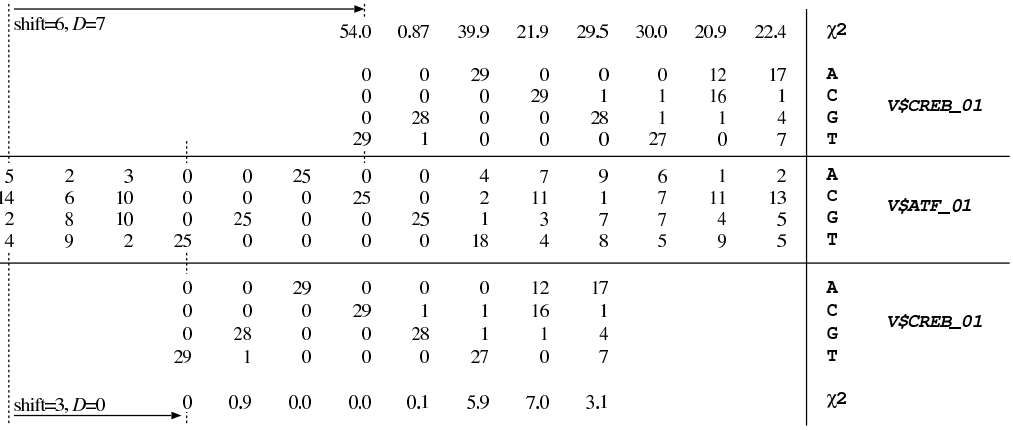


Figure 4.2: The distance D is computed for each possible shift between two matrices (CREB and ATF here). For all columns χ^2 values are calculated. D is then the number of χ^2 values exceeding the threshold $\chi^2_{\text{th}} = 7.81$. For shift= 6, the two matrices are not properly aligned, $D = 7$. For shift= 3, the two matrices are properly aligned, $D = 0$.

In order to keep the analysis easy, simply the number of significantly different positions within an overlapping part of the profiles is counted. The

example in Fig. 4.2 shows that for an appropriate alignment (with shift=3) of the two matrices all χ^2-values are below the χ^2_{th} threshold and hence no column appears to be different. Although the counts in some columns look quite different the limited sample size allows no statistically significant discrimination.

Obviously, the number of significantly different columns depends on the relative position of both matrices. All possible shifts with a minimum overlap of 6 bases and containing at least 75% of the information content of each matrix[1] are analysed. The minimal number of significantly different positions among these alignments D may be interpreted as the distance between the compared matrices. Fig. 4.2 illustrates that for a correct alignment of the ATF and CREB a distance $D = 0$ is obtained whereas other alignments lead to statistically significant different columns ($D > 0$).

An advantage of this distance measure in comparison to earlier studies [Hughes et al., 2000, Sandelin and Wasserman, 2004, Pietrokovski, 1996, Hannenhalli and Levy, 2002] is the emphasis on the limited sample size of many matrices. The expression 4.1 bases on counts of nucleotides observations, not only on their probabilities. Only few binding sites, such as those recognized by the Sp1 factor, are characterized by hundreds of experimentally verified sites. The more common sample size is around 15–20 (see Table 4.1) and, thus, it is much more difficult to distinguish matrices. The χ^2 measure leading to the distance D takes into account the limited sample size in a statistically well defined manner.

4.2.3 Correlation C of position weight matrices scores

The information on experimentally verified binding sites stored in position frequency matrices can be exploited to predict novel binding sites. For this purpose typically position weight matrices (PWMs) are constructed, which contain log-likelihood ratios to find a nucleotide at a position given an a PFM versus a random model. The PWMs are a natural way to score potential biniding of a matrix along an analysed DNA sequence. High positive scores appear around positions within the DNA sequence where binding of a factor is highly probable; negative scores mark places unprobable to be real binding sites. The similarity measure C is defined as the maximum correlation coefficient which can be observed between vectors of scores obtained for

[1]This constraint guarantees that the comparison is done over the positions carrying information. Otherwise, flanking parts of matrices corresponding to unspecific bases (`N`) are detected as similar.

two PWMs under study along a random test DNA sequence. In other words, C quantifies the intuitive observation, that when binding sites predicted by two weight matrices tend to overlap, then these matrices have to be similar to each other.

The position weight matrices (PWMs) can be constructed from the PFM counts $f_{b,i}$ in the following manner [Wasserman and Sandelin, 2004, Hertz and Stormo, 1999]. First, the probability $p_{b,i}$ of a base b at a given position i is estimated by:

$$p_{b,i} = \frac{f_{b,i} + s_b}{N_i + \sum_{b'=\mathtt{A,C,G,T}} s_{b'}},$$

where $N_i = \sum_{b'} f_{b',i}$ denotes the sample size at the position i leading to the relative frequency $\frac{f_{b,i}}{N_i}$. This estimator is modified using positive pseudo-counts s_b, which guarantee that the estimated probabilities are strictly positive even if zeros appear in the PFM. The influence of pseudo-counts s_b is strong when they are comparable to the counts f_b. The relatively large choice of $s_b = \frac{\sqrt{N_i}}{4}$ (suggested by Fickett [1996]), i.e. the pseudo-count being proportional to the standard deviation of the sample size, has a pronounced effect on PWMs with small sample sizes.

From the estimated probabilities $p_{b,i}$ the weights $w_{b,i}$ can be obtained as follows:

$$w_{b,i} = \log_2 \frac{p_{b,i}}{r_b},$$

where r_b refers to the *a priori* probability to find a base b in the DNA sequences. Consequently, the weights $w_{b,i}$ represent log-likelihood ratios to find a base pair b at a position i. When a nucleotide is overrepresented at a position of the binding site it gives a positive weight whereas underrepresented nucleotides are negatively weighted.

Finally, the score S_k around the position k of a test DNA sequence is a sum of the weights corresponding to nucleotides observed in the DNA sequence at the subsequent positions starting from the position k (see Fig. 4.3). The sum S_k is computed for each position k of the matrix along the DNA sequence. High positive scores S_k indicate locations in the test DNA sequence with strong binding affinities whereas zero or negative scores are found elsewhere.

This widely used technique of score calculation leads immediately to the correlation-based similarity measure (similar in spirit to the method used by Haverty et al. [2004], but modified to take into account the sample sizes of compared matrices). For two given matrices f and g one can directly obtain the corresponding score vectors S_k^f and S_k^g along all positions k of a given test DNA sequence. When the matrices f and g are highly similar positive score peaks are expected to occur at nearly the same positions k, i.e. prediction of nearly the same set of binding sites should be observed (see

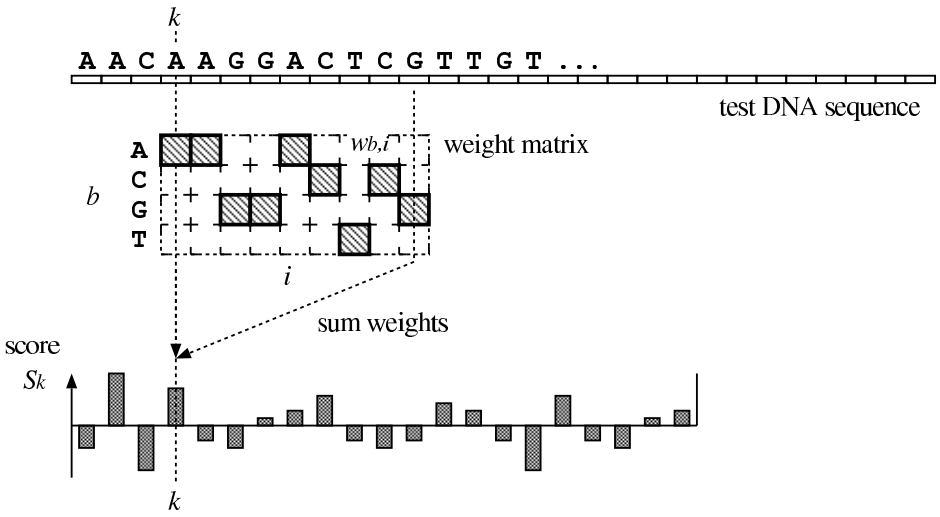


Figure 4.3: Calculation of the score S_k at the position k of the test DNA sequence for the weight matrix $w_{b,i}$.

Fig. 4.4). The Pearson correlation coefficient of the overlapping parts of the score vectors S_k^f and S_k^g provides a simple way to quantify the similarity of the predicted binding sites. Here also all possible relative shifts of both PWMs (corresponding to a minimum of 6 base pairs overlaps of the matrices) are considered. The similarity measure C is defined as the maximum correlation coefficient over all shifts.

4.3 Results

4.3.1 Comparison of both similarity measures

In the sections 4.2.2 and 4.2.3 two different measures of similarity of matrices, recognized by transcription factors, have been introduced. The first measure D is constructed around statistical properties of frequency matrices, representing a direct outcome of sequence alignment algorithms. The second measure C is based on weight matrices representation, present naturally in use-cases of the matrices. This chapter compares properties of both measures applied to the pairs of matrices listed in the Transfac database.

The first analysed question is, for which pairs of matrices a coincidence of small distances D and high correlation coefficients C is observed (i.e. for which matrices the two measures give consistent results). Fig. 4.5 shows three histograms of correlation coefficients C for matrices with small distances

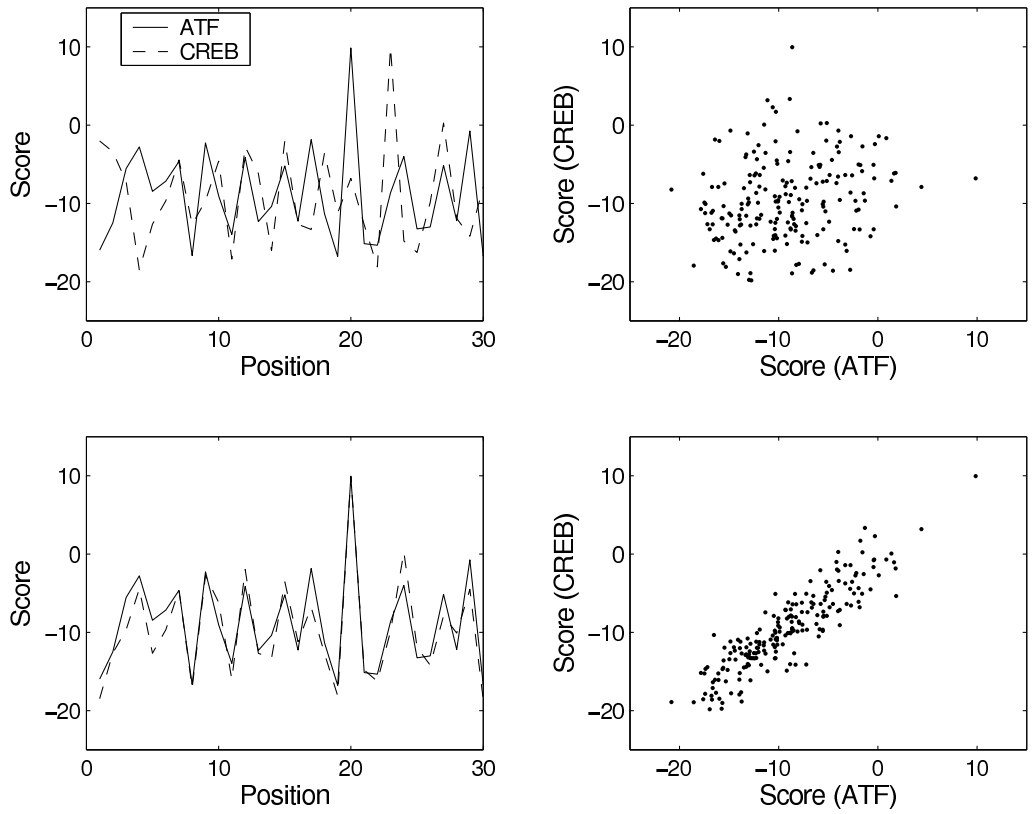


Figure 4.4: Correlations of ATF and CREB scores along a test DNA sequence. Left: first 30 scores for ATF (solid line) and CREB (dashed line). Right: scores for ATF versus scores for CREB. Only the first 200 scores are plotted, but the full length of the test DNA sequence is 10000 base pairs. Upper (shift=0): the matrices are not properly aligned ($C = 0.068$). Lower (shift=3): the matrices ATF and CREB are properly aligned and both reveal a binding site at position 20 ($C = 0.881$).

$D = 0, 1, 2$. As expected, there are many pairs of matrices with $D = 0$ and large values of C (see the right peak in the upper panel of Fig. 4.5). For such highly similar pairs both measures consistently find the same shifts (see Fig. 4.1) between corresponding matrices in the both similarity measures. Such matrices, although present in a database as different entries, can be treated as a single entry, since the differences between their positions are neglectable (based on the amount of observations used in alignment) and binding sites predicted with them are approximately the same.

However, there exists also many pairs of matrices with $D = 0$ and rela-

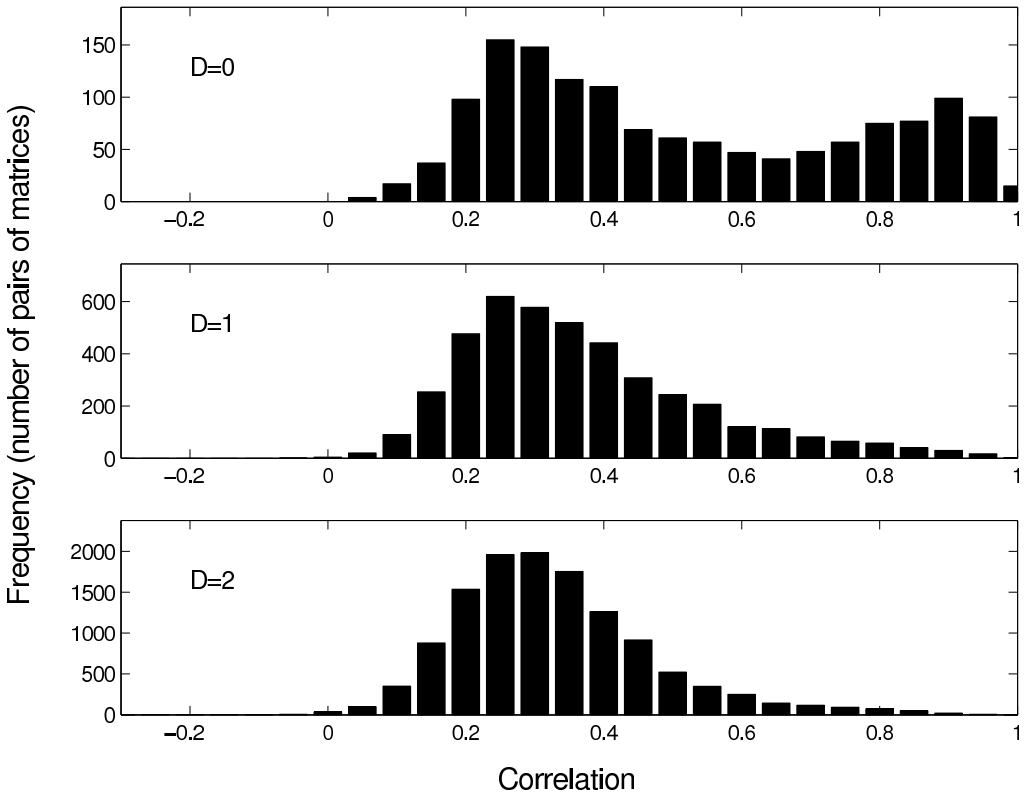


Figure 4.5: Histograms of the correlation C of the scores vectors obtained for small values of the distance $D = 0, 1, 2$ (number of significantly different columns according to the χ^2 test). These data have been calculated for pairs of matrices from the Transfac database. Please note: the vertical scales are different.

tively small correlation coefficients C (see the left peak in the upper panel of Fig. 4.5). These pairs refer mainly to matrices with a low information content and/or small sample size. In such cases the differences between columns are not statistically significant (many `N`s in both consensus sequences) but their scores along the test DNA sequence correlate only weakly. For example, the matrices `V$STAT4_01` and `V$MEF2_01` (see Transfac) are characterized by sample sizes $N = 6$, $N = 5$ respectively and have a distance $D = 0$ but a correlation only $C = 0.20$.

There are also cases with a high correlation coefficient C but with a distance $D > 2$. Such a situation appears for large matrices for which only a part is informative. For example matrices `V$GR_01` and `V$PR_01` (see Transfac) have a length of 27, but only six positions constitute the core sequence

(TGTTCT). Among the other positions three are significantly different, leading to a distance $D = 3$ but these differences affect the correlation C only weakly ($C = 0.92$).

Consequently, both introduced measures quantify independent properties and complement each other. In the following two potential measure applications are demonstrated: identifying redundancies and similarities in databases of matrices (see section 4.3.2) and checking novel matrices (see section 4.3.3).

4.3.2 Clusters of similar matrices in Jaspar and Transfac databases

Applications searching for binding sites typically observe scores of weight matrices provided by transcription factor binding site libraries. Obviously, redundancies present in such libraries can be directly observed in the search results. Such redundant results should be interpreted carefully, since they are a consequence of the library imperfection, rather than a biologically meaningful prediction. Therefore, similarities of matrices provided by the popular Jaspar and Transfac databases are studied in this section.

As it has been demonstrated in section 4.3.1 the two measures D and C catch different and independent features of matrices, so both measures are used in this study. Here, pairs of matrices for which $D \leq 1$ and $C \geq 0.8$ as considered similar (see Fig. 4.5). These constraints imply that the matrices are almost indistinguishable from a statistical point of view (D close to zero) and that their scores along DNA sequences are strongly correlated (C close to one).

Fig. 4.6 provides an overview of comparison of Jaspar (ellipses) and Transfac (boxes) matrices. Even though the details of the figure are only readable in the online version, the presence of many lines connecting similar matrices is visible. Consequently, the used technique allows an automatic "alignment" of these collections of matrices. This is not a trivial task since the naming conventions used in the databases are different, and thus finding matrices corresponding to each other requires expert knowledge. A detailed list of 84 Jaspar matrices having highly similar counterparts (with $D \leq 1$ and $C \geq 0.8$) in the Transfac database is provided in Tab. B.1. Another 16 Jaspar matrices have lower similarities $D \leq 3$ and $C \geq 0.6$. Only the Jaspar matrices P_HMG-1, P_HMG-IY and V_Ghlf, have no obvious "partners" in Transfac.

In addition to the similarities of Transfac and Jaspar matrices, there exist also similarities of matrices originating from the same library. Clusters

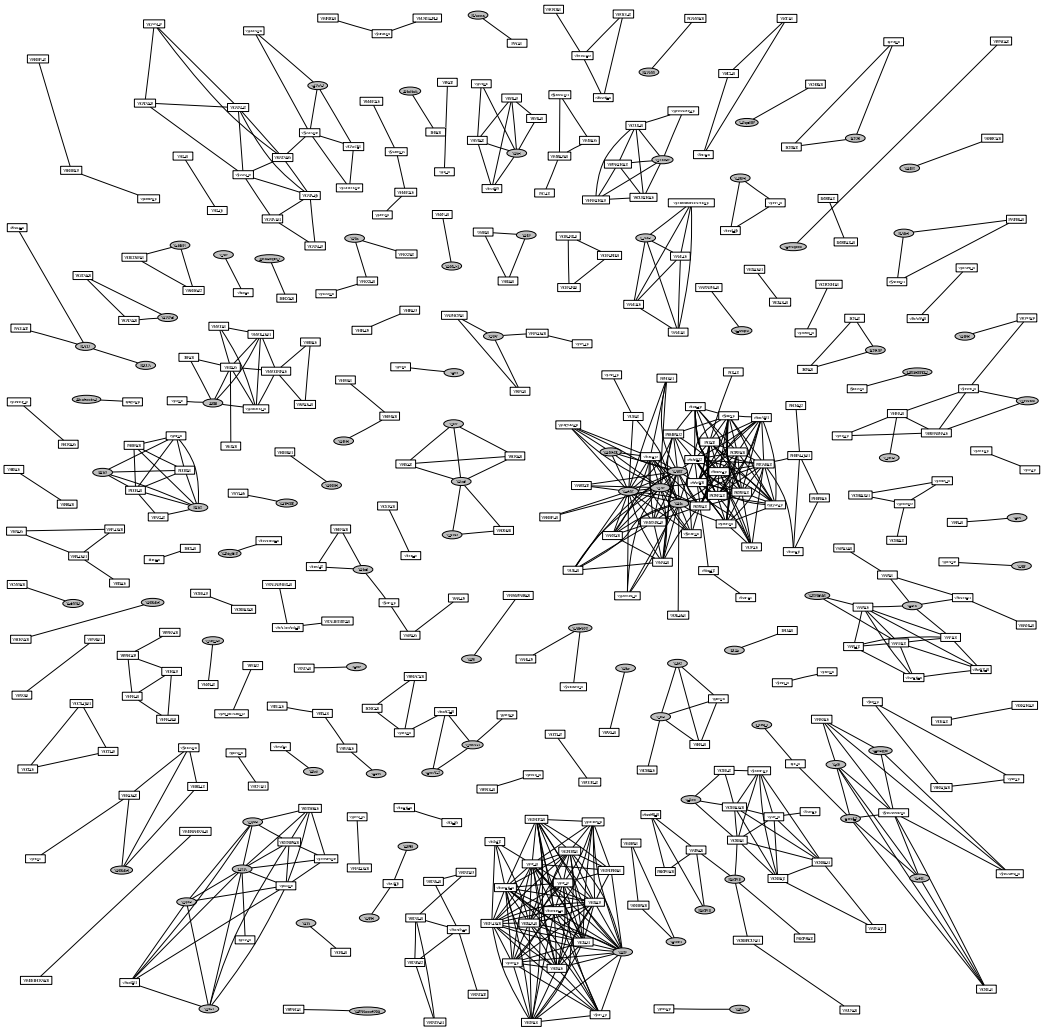


Figure 4.6: Similar matrices of Jaspar (gray ellipses) and Transfac (white boxes) databases. A connection is drawn, when the similarity measures $D \leq 1$ and $C \geq 0.8$. Detailed list of shown matrices is provided in Tab. B.1.

of such matrices reflect pronounced redundancies in the matrix collections. These redundancies are mainly inherent in the construction of the databases. There are for example, matrices of the same transcription factor with different degrees of stringency (see for instance AP1 matrices). Moreover, different transcription factors of certain families have almost identical binding motifs (see for example Myc-Max, USF and ARNT). A complete list of all clusters is provided in Tab. B.2. An interesting collection of structural classes of transcription factors has been compiled recently by Sandelin and Wasserman [2004]. Consistent with their results, here also clusters of the ETS family (see cluster 2 in Tab. B.2, also enlarged in Fig. 4.7(b)), bHLH transcription factors (cluster 15), and REL family (cluster 5) are identified.

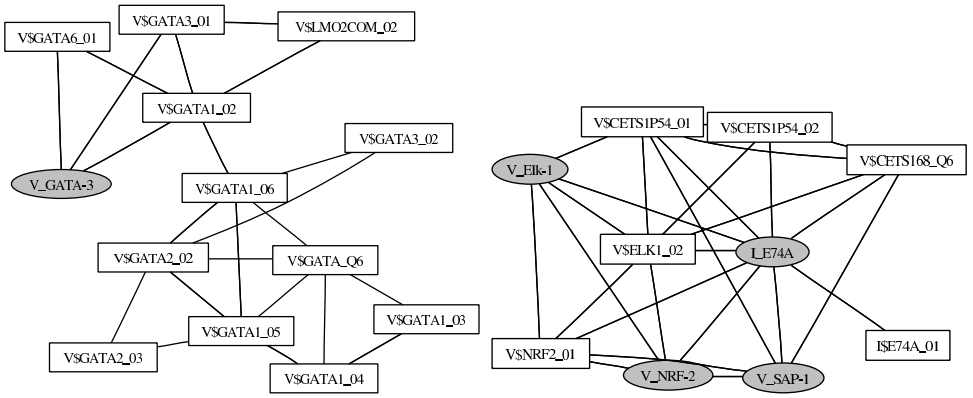


Figure 4.7: Transcription factor families (a) GATA and (b) ETS.

In Fig. 4.7 enlargements of two selected clusters representing the GATA and ETS transcription factors families are shown. The high similarity of these matrices cannot be directly noticed by inspection of names or consensus sequences. Furthermore, subgroups might be detected using the proposed statistical approach. For example, the GATA cluster reveals that the Jaspar matrix has particularly high similarity to the Transfac entries `GATA1_02`, `GATA3_01` and `GATA6_01`, but less similarities to other members of the GATA class.

The redundancies visualized in Figs. 4.6 and 4.7 can be exploited to reduce the number of matrices. Highly similar matrices match a DNA sequence either both or not at all. Therefore, one could construct "consensus matrices" as in Sandelin and Wasserman [2004] or one might select representative matrices in each cluster. In this way the number of false predictions in the search for transcription factor binding sites can be decreased [Sandelin and

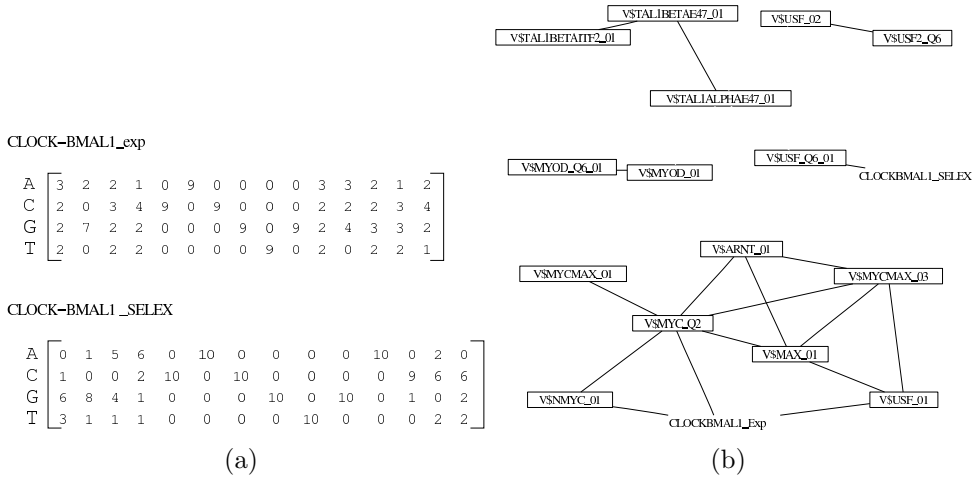


Figure 4.8: (a) CLOCK-BMAL1 matrices based on experimentally characterized binding sites of clock genes and from a SELEX study. (b) Mapping of CLOCK-BMAL1 matrices on E-box matrices. These matrices have been selected from the Transfac database and include MYC, MAX, ARNT, MYOD, USF, TAL1/E47 (see Munoz et al. [2002] for a review on E-box transcription factors). An edge is drawn when $D \leq 1$ and $C \geq 0.8$.

Wasserman, 2004].

4.3.3 Mapping of novel matrices

A careful inspection of the clusters found automatically by the similarity analysis in section 4.3.2 might reveal unexpected similarities pointing to possible cross-talks of different signaling cascades on the level of transcriptional regulation. One possible example might be the regulation of circadian clock genes and cell cycle control [Cardone and Sassone-Corsi, 2003, Matsuo et al., 2003]. In both processes bHLH transcription factors bind as dimers to E-boxes. The corresponding Myc-Max cluster appeared already in as the largest cluster in Fig. 4.6. In the mammalian circadian clock the CLOCK-BMAL1 dimer regulates clock genes such as *Per1*, *Per2*, *Per3*, *Cry1* and *Cry2*. A matrix describing explicitly the binding sites of CLOCK-BMAL1 could not be found neither in Jaspar, nor in Transfac. Consequently, such a matrix might be constructed in the following ways. On one hand 9 experimentally verified binding sites from 7 different clock genes are available [Darlington et al., 1998, Gekakis et al., 1998, Chen and Baler, 2000, Munoz et al., 2002, Vallone et al., 2004]. On the other hand, 10 sequences with high affinities to the

CLOCK-BMAL1 dimer have been found in a SELEX experiment [Hogenesch et al., 1998].

Both matrices are visualized in Fig. 4.8(a). They contain the E-box consensus motif `CACGTG`, but differ in the flanking regions. Details of the matrix construction are given in Kielbasa et al. [2005].

Fig. 4.8(b) shows that these novel matrices have highly similar counterparts in Transfac (NMYC, MYC, USF). Consequently, cross-talk of the circadian clock with cell cycle regulation and tumor genesis can be expected at the level of transcriptional control. Indeed, the success of chronotherapies and recent detailed studies on cross-talk underline the dependence of circadian rhythms with tumor growth [Filipski et al., 2002]. Also in the process of liver regeneration a pronounced effect of the circadian clock on cell cycle control has been found [Schibler, 2003]. This example illustrates that a careful SELEX experiment combined with a mapping of the resulting matrix to known matrices can reveal possible functions of the corresponding transcription factor.

4.4 Discussion

Understanding gene regulation in higher eukaryotes is still challenging and current computational algorithms suffer from a large amount of false positive predictions [Wasserman and Sandelin, 2004, Kielbasa et al., 2004b]. In particular the large number of position frequency matrices in databases such as Jaspar or Transfac leads to high-scoring predictions every few base pairs. Consequently, a careful pre-selection of matrices is essential. On one hand, expert knowledge can be used to select a subset of candidate matrices for the analysis of upstream regions (illustrated also in section 3.3). Such a selection is, however, subjective and novel combinations of transcription factor binding sites might be missed. On the other hand, for large scale computational studies, it is useful to have an automatic tool to detect similar matrices. Therefore, in this chapter two independent similarity measures of matrices representing profiles recognized by transcription factors are studied. This measures are used to quantify redundancies in the databases, to map the entries of different databases, and to cluster matrices.

The first similarity measure, defined in section 4.2.2, is based on a χ^2 test. In contrast to earlier approaches based on normalized frequencies [Hughes et al., 2000, Sandelin and Wasserman, 2004, Pietrokovski, 1996] additionally the small sample size of many matrices is taken into account. The number of significantly different matrix columns of two position frequency matrices is counted and it defines the distance D of the two matrices. This chapter

focuses on highly similar matrices with $D \leq 1$, although in forthcoming studies the χ^2 measure might be taken directly to calculate distances of matrices in more detail.

The second similarity measure, specified in section 4.2.3, is related to the primary application of position weight matrices – the prediction of binding sites in uncharacterized DNA sequences. For two matrices of interest the match scores along a test DNA sequence are calculated and the Pearson correlation coefficient of the corresponding score vectors constitutes the measure C. Thus large values of C indicate that both matrices predict essentially the same binding sites. In this chapter a long random nucleotide sequence was chosen as the test DNA sequence. However, the measure can be easily adapted also to other test sequences such as sets of promoter regions.

A combination of both similarity measures was first used to map the Jaspar matrices to the Transfac database automatically (see section 4.3.2). Requiring rather strong similarity ($D \leq 1$, $C \geq 0.8$), redundancies in these databases were quantified and clusters of almost indistinguishable matrices were constructed (see Tables B.1 and B.2). This allows to reduce the number of matrices used in search for transcription factor binding sites. Since the number of non-redundant matrices is still high, the number of false positives remains a problem and multiple testing correction is required. Consequently, further approaches such as phylogenetic footprinting [Wasserman and Sandelin, 2004, Dieterich et al., 2002, Ureta-Vidal et al., 2003], transcriptional profiling [Bussemaker et al., 2001], or ChIP on chip experiments [Ren et al., 2000, Martone et al., 2003] have to be combined with a preselection of non-redundant matrices. Another independent method utilizing gene annotations, TFGossip, which can also be used to improve quality of binding site predictions is presented in chapter 5.

The proposed measures can be also used to predict cross-talk on the level of transcriptional control. As an illustration the cluster of E-box binding bHLH transcription factors is discussed (section 4.3.3). Since circadian clock genes are regulated by a binding site quite similar to the Myc-Max motif, a strong interdependence of circadian regulation and cell cycle control is expected and is indeed known empirically for decades in connection with chronotherapies or liver regeneration. Finally the similarity measures are used to assign newly derived matrices to known transcription factors. An E-box matrix obtained from SELEX experiments with the CLOCK-BMAL1 dimer is mapped to the Myc-Max cluster as an illustration. Thus the possible function of poorly characterized transcription factors can be predicted using affinity measurements (SELEX) combined with a comparison of the resulting matrix to database matrices.

Chapter 5

Prediction of functions of transcription factors (TFGossip)

Summary

In this chapter a functional view on the *in silico* prediction of transcriptional regulation is proposed. The study presented here has been published in Kielbasa et al. [2004a] and Blüthgen et al. [2005b] as an outcome of a joined project of Nils Blüthgen and Szymon M. Kiełbasa contributing to it in equal parts. Section 5.2.1 introduces TFGossip – a method for predicting biological functions regulated by combinatorial interactions of transcription factors. Using a rigorous statistic this approach intersects the predictions of transcription factor binding sites in gene upstream sequences with Gene Ontology terms associated with these genes. Positional frequency matrices describing the profiles recognized by transcription factors constitute the input of the method and significantly enriched biological processes are reported as the output.

The proposed algorithm was tested carefully on several sets of matrices. Section 5.3.1 presents terms predicted to be associated with factors of cell cycle related E2F family. Clustered occurrences of predicted binding sites corresponding to a single E2F matrix are studied there. Section 5.3.2 discusses a pair of transcription factors NFAT/AP-1 involved in immune response. The last example for a well-studied set of skeletal muscle related transcription factors Myf-2, Mef and TEF is given in section 5.3.3. In all these cases the reported results match well the experimental knowledge. Furthermore, for the NFAT/AP-1 composite element, as well as for the muscle related factors

novel functions are predicted.

5.1 Introduction

The regulation of transcription is a major mechanism controlling the spatial and temporal activity of genes, thereby governing the organization of biological processes in eukaryotic organisms. A complex signaling machinery transduces external and internal stimuli to the activities of transcription factors which are the major means of transcriptional regulation. Through this, eukaryotic cells are equipped to adapt adequately to the environment and to orchestrate events like proliferation and differentiation. In contrast to prokaryotes, where transcriptional regulation can be understood in terms of induction by single factors, the regulation in eukaryotes is mainly carried out by sophisticated interactions of multiple transcription factors (for an example, see Yuh et al. [1998]). Additionally, the regulatory sites are distributed over large regions of the genome including intronic sequences [Cawley et al., 2004, Euskirchen et al., 2004]. It is rare that individual binding sites are strongly conserved, only the combinatorial action gives rise to a specific control. Therefore, understanding complex gene regulatory networks in higher organisms is an extremely difficult task.

Considering the importance of transcriptional regulation and the vast amount of genomic data available, automated inference of the gene regulatory network is a major challenge in the post-genomic era. However, a straight forward search for transcription factor binding sites represented by consensus sequences or weight matrices leads to the curse of false positives. Wasserman and Sandelin [2004] estimate that a simple search for binding sites results in only one functional site per 1000 predictions. Consequently, other available biological properties of gene regulation have to be exploited to improve computational predictions. For instance, groups of co-regulated genes from expression profiling [Pilpel et al., 2001, Kielbasa et al., 2004b], phylogenetically conserved regions [Wasserman et al., 2000, Dieterich et al., 2003] and the clustering of binding sites [Frith et al., 2003] are studied. Nevertheless, the specificity of binding site prediction is still unsatisfactory [Wasserman and Sandelin, 2004] and expensive experimental studies such as ChIP on chip experiments [Ren et al., 2000, Martone et al., 2003, Euskirchen et al., 2004] are necessary.

In this chapter a new approach, TFGossip, is described proposing a functional view on the gene regulatory network by utilizing the growing systematic representation of expert knowledge compiled in the Gene Ontology [Ashburner et al., 2000]. The presented algorithm uses public software to ex-

tract upstream regions of genes [Blüthgen et al., 2004] and to predict clusters of binding sites [Frith et al., 2003]. Then the genes with a common cluster in their upstream regions are searched for statistical association with annotations from the Gene Ontology. For this purpose a novel program Gossip [Blüthgen et al., 2005a] is used, which precisely takes into account multiple sample correction issue. This approach allows prediction of biological functions controlled by combinatorial action of transcription factors. It has been demonstrated [Kielbasa et al., 2004a] that this approach works for predicting the functional regulation of both a single factor (section 5.3.1) and a combination of two well-characterized factors (section 5.3.2). The inference of more complex, combinatorial regulation by several transcription factors was successfully verified [Blüthgen et al., 2005b] based on a well-studied set of transcription factors that co-regulate skeletal muscle gene expression [Wasserman and Fickett, 1998]. It turns out that without using *a priori* knowledge about the functional targets of these factors we can predict their function correctly (section 5.3.3). Subsequently it has been shown, that the proposed approach can also bridge the gap between detailed studies of the regulation of single genes and genome-wide analysis. Fessele et al. [2001] have unveiled the transcription factors that differentially regulate the expression of the chemokine RANTES upon lipopolysaccharide stimulation in monocytes. The TFGossip framework was applied to this set of transcription factors and the predicted functions were compared with profiles of publicly available microarray data Blüthgen et al. [2005b]. The results show a remarkable similarity, although the microarray data have been generated in a different organism and after a long stimulation, allowing also indirect regulation.

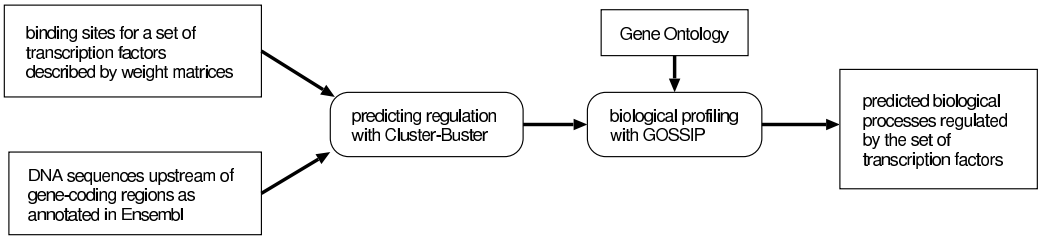


Figure 5.1: Data flow in TFGossip: Clusters of binding sites are searched in gene upstream regions using the Cluster-Buster algorithm. The list of genes having a cluster predicted is then passed to Gossip, which detects association of the genes with biological processes using the Gene Ontology. The significantly associated processes are reported.

5.2 Materials and Methods

5.2.1 TFGossip

The analysis presented in this chapter combines two algorithms (see Fig. 5.1). First, a genome-wide search of the genes that are potentially regulated by the transcription factors under consideration is performed. For this purpose, the Cluster-Buster program [Frith et al., 2003] is applied to predict clusters of transcription factor binding sites in upstream regions of genes (reference group). To avoid problems arising from subjective parameter tuning, the default parameters of the program are used.

Second, the genes with a cluster predicted are tested for association with biological processes. This is performed by Gossip (section 5.2.2) using Gene Ontology annotations [Ashburner et al., 2000]. This algorithm tests each term in the Gene Ontology for enrichment in the annotations of genes from a test group compared to those from a reference group. Here, the test group contains the genes for which Cluster-Buster reports a cluster of binding sites. The reference group is composed of all genes under study. Fig. 5.2 illustrates the TFGossip algorithm in details.

5.2.2 Gossip

This section introduces shortly the Gossip algorithm as described in Blüthgen et al. [2005a]. Using the one-sided Fisher's exact test, which is based on the hypergeometric distribution, Gossip calculates a p-value, for the null hypothesis that the annotations for the test group are sampled randomly from the reference group. Since this test is performed on all Gene Ontology terms,

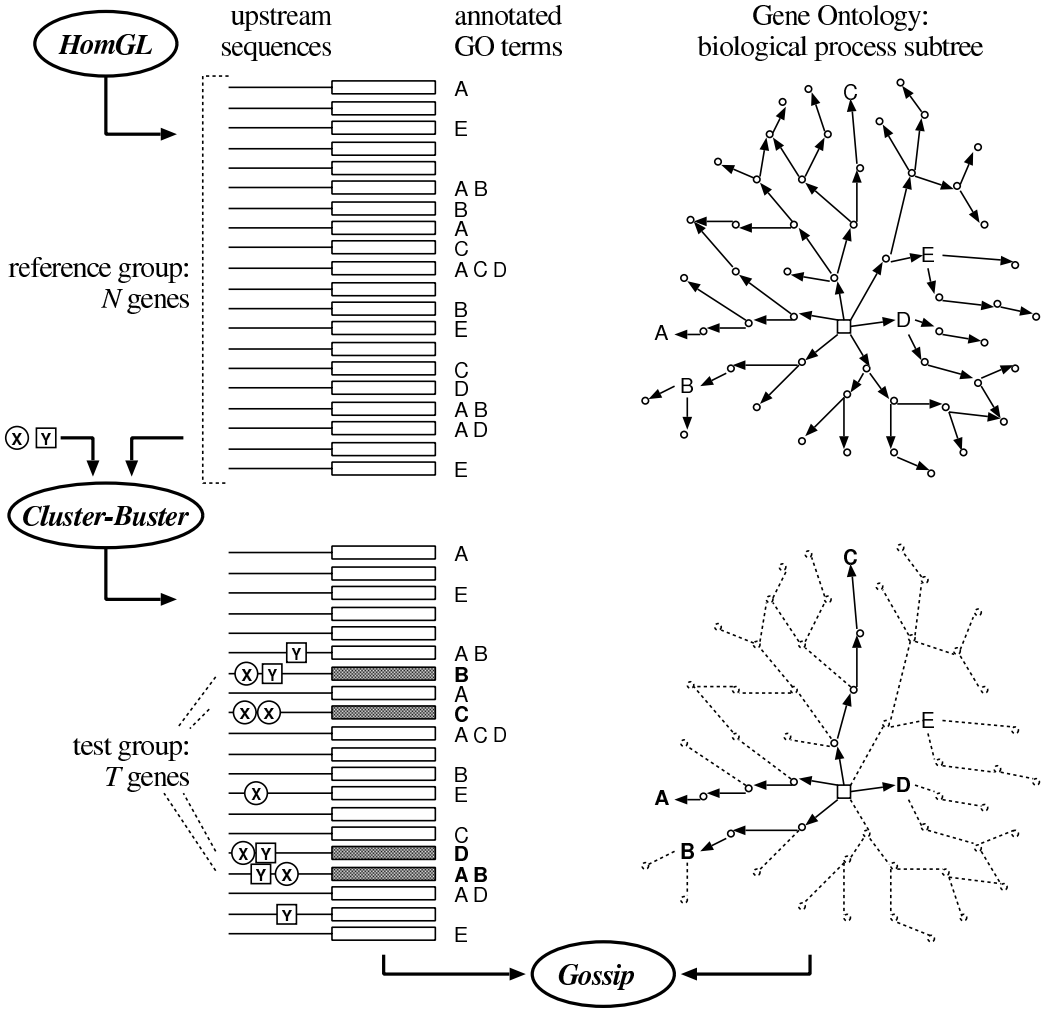


Figure 5.2: Details of the TFGossip algorithm. HomGL extracts upstream sequences and Gene Ontology annotations (e.g. A,. . . , E) for all N human reference genes. Cluster-Buster selects a test group of T genes with clusters of predicted binding sites of user defined transcription factors (X, Y). The terms annotating the test genes, as well as their parent terms in Gene Ontology hierarchy, are potential candidates to be overrepresented in the test set given the reference set. All these terms are tested by the Gossip algorithm, and the significantly overrepresented ones are reported.

problems arising from multiple testing have to be taken into account. Therefore, the authors decided not to use single-test p-values but the False Discovery Rate (FDR) is taken as an adequate measure of significance. FDR(α) quantifies the expected Number of False Discoveries $\langle \mathrm{NFD}(\alpha) \rangle$ in relation to the total Number of Positives NP(α) at a single test p-value threshold α:

$$\mathrm{FDR}(\alpha) = \frac{\langle \mathrm{NFD}(\alpha) \rangle}{\mathrm{NP}(\alpha)} \ .$$

Using the hypergeometric distribution, the expected number of false discoveries $\langle \mathrm{NFD}(\alpha) \rangle$ can be calculated for each p-value threshold α by

$$\langle \mathrm{NFD}(\alpha) \rangle = \sum_i \sum_j^{p_f(j,Z_i,T,N)<\alpha} h(j, Z_i, T, N) \ .$$

Here the first sum runs over all terms i from the Gene Ontology, and in the second sum the probability of j genes being annotated with term i is summed up as long as the one-sided Fisher's exact test $p_f(j, Z_i, T, N)$ does not exceed α. Z_i denotes the number of genes annotated with the term i in the reference group. T and N denote the number of genes in the test group and in the reference group, respectively. $h(j, Z_i, T, N)$ represents the hypergeometric distribution:

$$h(j, Z_i, T, N) = \frac{Z_i! T! (N - Z_i)! (N - T)!}{j! N! (Z_i - j)! (T - j)! (N + j - Z_i - T)!} \ .$$

Within this chapter, such a threshold α is chosen, that the false discovery rate is kept below 5%. Further details and the Gossip software are available at the website `http://itb.biologie.hu-berlin.de/~nils/gossip`.

5.2.3 Data preparation

Gene sequences analysed in this chapter correspond to 16,032 human UniGene [Wheeler et al., 2003] clusters and were extracted upstream of the transcription start sites reported by Ensembl [Birney et al., 2004]. Out of them 15,362 unique upstream regions were identified, since several UniGene clusters pointed to the same genes in Ensembl. The duplicates were treated as single genes and their Gene Ontology annotations were composed together. Sequences of lengths 250, 500, 750, 1000, 1250, 1500, 2000 bp upstream of the TSS were tested in the analyses, although it has been found that typically using length of 1000 bp showed terms with the lowest p-values (see Blüthgen et al. [2005b], Supplementary Fig. S2). This is in agreement with

the estimate by Dieterich et al. [2002], that the majority of promoters should overlap with these regions.

The Gene Ontology [Ashburner et al., 2000] defines a hierarchical controlled vocabulary to annotate genes. It contains three main branches: *biological process*, *molecular function*, *cellular location*. Each annotation implies a series of more general annotations upward in the hierarchy of the Gene Ontology, which are also taken into account. The analysis presented here was limited to the branch describing biological processes. The annotations from the Gene Ontology were assigned to the genes using HomGL [Blüthgen et al., 2004].

5.3 Results

5.3.1 Functions of E2F transcription factor

E2F is a transcription factor family known to be involved in the regulation of the S-phase of the mitotic cell cycle [Herwig and Strauss, 1997]. A profile recognized by members of this family is provided by the Transfac database [Heinemeyer et al., 1998, Matys et al., 2003] at the accession number `M00427` (`V$E2F_Q6`). This transcription factor was chosen to test whether the TFGossip pipeline is able to predict functions regulated by members of a single family of transcription factors, acting in a clustered way. The Cluster-Buster program predicted 273 genes to contain a cluster of binding sites matching the E2F profile. In these 273 target genes, many biological processes related to the *cell cycle* are significantly enriched (see Fig. 5.3). The more specific processes are related to *DNA replication* and *S-phase of the mitotic cell cycle*. Therefore the prediction of the TFGossip pipeline yields exactly what is known from experiments: E2F is regulating the synthesis phase of the cell cycle.

5.3.2 Functions of NFAT and AP-1 transcription factors

Kel et al. [1999] have studied the composite binding of NFAT and AP-1 transcription factors. NFAT plays a role in the regulation of cytokines and other genes during immune response. NFAT co-operates with AP-1 to integrate calcium and PKC-signal transduction. The authors concluded that the combination of NFAT and AP-1 exhibits a significantly higher specificity than individual factors towards genes induced upon T-cell activation.

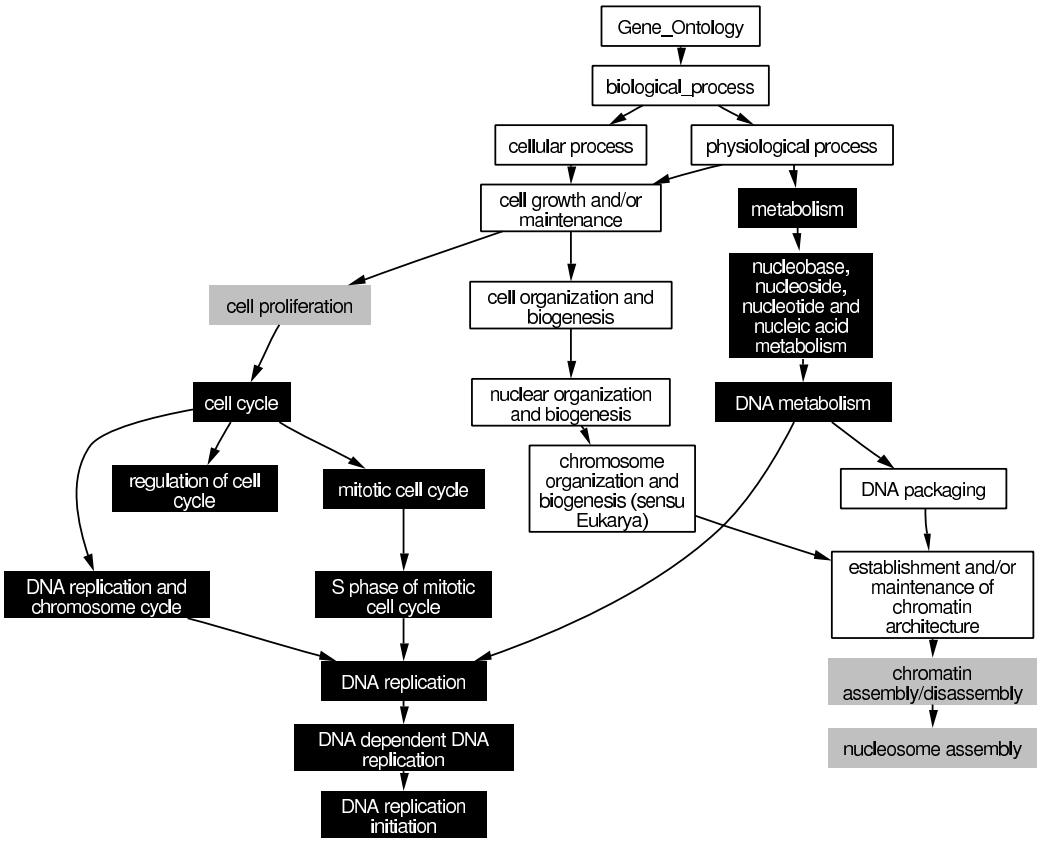


Figure 5.3: A part of the Gene Ontology highlighting the biological processes that are significantly enriched in 273 genes with a cluster of E2F-binding sites predicted in their upstream regions. Black boxes show those processes with FDR$\leq$ 0.01, gray boxes with FDR$\leq$ 0.05.

The TFGossip pipeline was applied to each of the corresponding matrices separately and no significantly enriched process was found within the predicted targets of individual factors (NFAT: 2716 predicted target genes – no process significant, AP-1: 921 predicted target genes – no process significant). However, studying both factors together revealed significantly enriched biological processes within the predicted 1913 target genes (see Fig. 5.4). Many of these overrepresented biological processes are related to immune response: *inflammatory response*, *response to wounding*, *innate immune response*, *response to biotic stimulus* which is in good agreement with the results of Kel et al. [1999]. Moreover the processes *organogenesis*, *morphogenesis*, *development*, *regulation of cell growth*, *regulation of cellular process*, and *cell adhesion* found as novel candidates might be regulated by the combinatorial

action of NFAT and AP-1.

5.3.3 Processes regulated by muscle transcription factors

Wasserman and Fickett [1998] have studied transcription factor families associated with skeletal muscle specific gene expression. Having identified transcription factors regulating skeletal muscle specific genes, the authors have constructed a set containing five positional frequency matrices based on binding sites (Mef-2, Myf, SRF) selected *in vitro* or on promoters, that do not play any role in the muscle-specific expression (Sp-1, TEF). The promoters of the skeletal muscle specific genes were intentionally not used to generate the matrices. Applying the TFGossip algorithm to all five matrices yields significantly associated biological processes: *muscle development* (FDR=0.003), *muscle contraction* (FDR=0.008) and *B-cell activation* (FDR=0.017). Since the Sp-1 factor is involved in the regulation of many other functions, a detailed study with all 26 matrix subsets containing at least two matrices out of the original five matrices was performed. In 12 cases no Gene Ontology term was found significantly overrepresented. The set consisting of Mef-2, Myf and TEF matrices was the most specific for *muscle contraction*. Also the term *striated muscle contraction* was reported with the highest significance here. The results for these matrices are shown in Figs. 5.5 and 5.6. Interestingly, *smooth muscle contraction* was not significant in any of the analyzed sets. This finding independently confirms, that the selected transcription factors may contribute to skeletal muscle-specific expression [Wasserman and Fickett, 1998]. Detailed results for all 26 subsets are listed in Supplementary Tab. S1 of Blüthgen et al. [2005b].

Most subsets containing the Mef-2 matrix resulted in terms related to *B-cell activation*. Transcription factors of the Mef-2 family are differentially expressed in B-cells and these cells have Mef-2C-containing, Mef-2-specific DNA binding complexes, suggesting a possible role for Mef-2C activity in B-cells [Swanson et al., 1998].

5.4 Discussion

The availability of whole-genome sequences and the growing systematic annotations like the Gene Ontology provide the means for more function oriented data mining beyond the level of single genes. TFGossip offers a method which allows the inference of biological functions regulated by a combinatorial interaction of transcription factors *in silico*. Contrary to other widespread

techniques this method does not intend to predict which factors control genes of similar expression profiles. Instead it only requires a set of positional frequency matrices representing profiles recognized by transcription factors to predict their biological function. First, using Cluster-Buster algorithm a list of potential target genes for a set of transcription factors is predicted (section 5.2.1). Afterwards a rigorous statistical test for association with biological processes implemented in Gossip (section 5.2.2) is applied to all biological processes provided by the Gene Ontology. Therefore the search is not biased by any prior knowledge related to the factors and gives a chance to detect novel regulatory associations.

The method's utility is demonstrated on several well studied examples. When only one matrix is provided, clusters of binding sites predicted with that matrix are analysed. The study of a single matrix corresponding to the E2F transcription factors (section 5.3.1) shows a successful application of the method. Moreover, as presented in section 5.3.2, functions regulated by two transcription factors binding cooperatively can be also detected. Finally, the most complex case of five transcription factor families involved in skeletal muscle specific gene expression (section 5.3.3) results with significantly enriched terms that are specific for skeletal muscle. Furthermore, as a novel prediction, the transcription factors Mef-2, SRF and a member of the TEF family are suggested to act together in the context of B-cell activation. Notably, no tuning of parameters was necessary within these studies.

In addition, in Blüthgen et al. [2005b] another application of TFGossip is discussed. Fessele et al. [2001, 2002] investigated transcription factors which bind the RANTES/CCL5 promoter in human monocytes differently in untreated and lipopolysaccharide (LPS) stimulated cells. TFGossip analysis for this set of factors (AP1, CEBP, CREB, ETS, NF-κB (p50 and p65), and Sp-1) reveals terms describing the biological processes regulated in monocytes after being exposed to LPS. The biological validity of the result is evaluated based on a microarray data set for LPS-stimulated mouse monocytes[1]. Here, Gossip was used to predict the biological processes associated with the up-regulated genes. Both lists of terms show remarkable agreement although they were predicted using independent sources of data related to different spicies. Moreover, TFGossip provided a sublist of all genes present on the microarray, which had both a predicted cluster of binding sites in their upstream regions and an annotation with at least one of the predicted terms. Then, based on the microarray data the fractions of up-regulated genes were

[1]generated by the Alliance for Cellular Signaling, `http://www.signaling-gateway.org/data/micro/cgi-bin/micro.cgi?expt=operon`, accession numbers: MAE040216Z53, MAE040217Z53, MAE040218Z53, MAE040216Z63, MAE040217Z63 and MAE040218Z63

counted: in the whole microarray gene list, and in the TFGossip sublist. For several fold change thresholds (1.5, 2.0, 2.5, 3.0, and 3.5) it has been observed, that the TFGossip-filtered list of genes contained approximately two times larger fraction of up-regulated genes than the whole microarray. Concluding, usage of functional annotations improves the specificity of genome-wide identification of transcription factor binding sites. Particularly interesting is the fact, that the TFGossip prediction has been performed on human sequences and the experiment has been done in mouse monocytes, which reflects a high degree of evolutionary conservation.

To assess the specificity of TFGossip, the predictions were compared to results obtained from random sets of positional frequency matrices. These random sets were constructed by permuting the positions of the matrices, thereby preserving their information content and GC-content. The analysis of 1200 sets of permuted muscle-related matrices Mef-2, Myf and TEF yielded 76 sets (6.3%) with one or more significant terms. Interestingly, none of the permuted data sets yielded results related to *muscle development* or *B-cell activation*. On average, we found 0.46 false discoveries with $p \leq 0.0008$ (the p-value of the least significant term *lymphocyte differentiation*). Considering the eight significant terms in the original data set (see Fig. 5.5), this corresponds to a false discovery rate of 5.8%.

The lack of functional data for combinatorial gene regulation in higher eukaryotes makes it difficult to construct a true positive set, and, consequently, to estimate the sensitivity of TFGossip. However, there are promoters with clusters of binding sites recognized by the same factor (for example clusters of E-boxes found in the promoters of circadian clock genes [Gekakis et al., 1998], or clusters of E2F binding sites which can clearly be associated with the S-phase of the cell-cycle). These clusters can be specific enough to unveil their functional targets. Therefore, profiling of single factors might provide a rough estimate of the sensitivity. TFGossip was applied to the 78 matrices for mammalian transcription factors from the Jaspar library [Sandelin and Wasserman, 2004] and the human upstream regions. 20 significant functional profiles with 142 terms in total were identified. Since the true functions of the factors were not known, the number of false profiles was estimated by permutation analysis. Here, on average 0.9 profiles with 4.3 terms were found. Given these numbers it can be estimated that about 19 factors out of the 78 factors under study can be correctly associated with their functional targets. Since it is not known which fraction of the 78 transcription factors exhibit clusters of binding sites in their target genes, this number cannot be translated directly into a specificity.

Finally it should be noted, that care must be taken in interpreting the outcome of TFGossip, when the factors given at the input recognize sites with

high GC-content (Sp-1, NF-κB p50). Although Cluster-Buster uses a background model that takes GC-variation into account, such factors prefer sites in GC-rich upstream regions. This affects the analysis since these regions themselves are associated with certain processes. The 10% of genes having the highest GC-content in their upstream regions are significantly associated with 35 terms describing processes like *development/neurogenesis*, *regulation of transcription*, *ion transport*, *phosphorylation* and *signal transduction*. Therefore, for a correct interpretation, the composition of the positional frequency matrices must be taken into account, as GC-rich matrices can induce more false positives. In such cases an analysis of permuted matrices can be used to find the expected number of false discoveries.

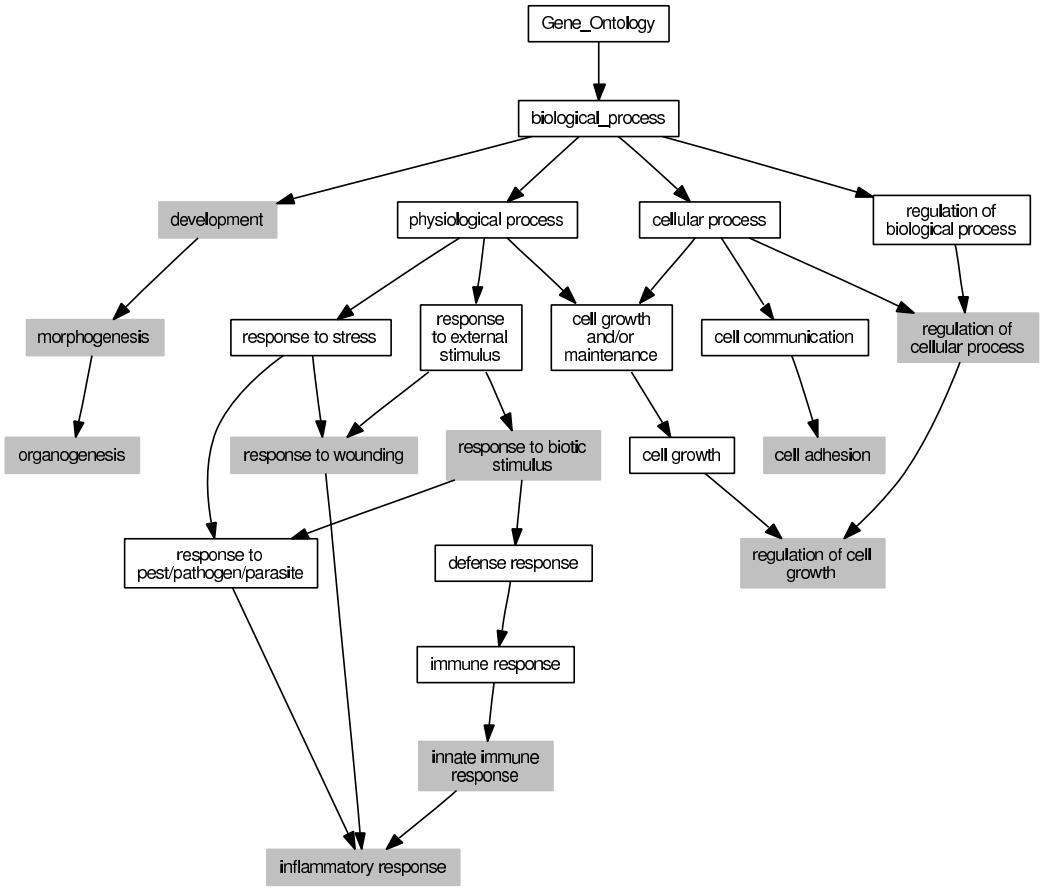


Figure 5.4: A part of the Gene Ontology highlighting the biological processes that are significantly enriched in 1913 genes with a cluster of NFAT and AP-1 binding sites predicted in their upstream regions. Grey boxes mark those processes with False Discovery Rate below 0.05.

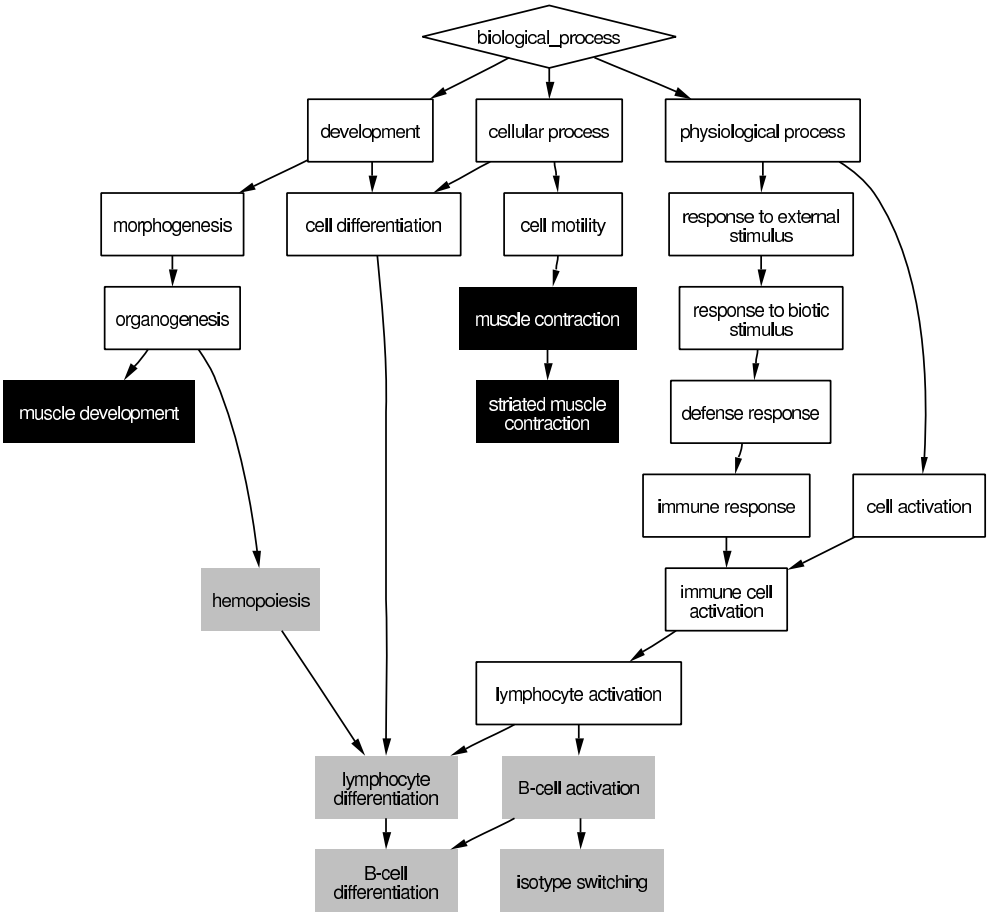


Figure 5.5: Significantly overrepresented Gene Ontology terms associated with binding sites of the transcription factors Mef-2, Myf and TEF, known to regulate the expression of muscle-specific genes [Wasserman and Fickett, 1998]. The black and grey boxes correspond to significantly overrepresented biological processes of the Gene Ontology within the predicted target genes (thresholds of FDR ≤ 0.01 and FDR ≤ 0.05, respectively). The diamond shows the root node for biological processes.

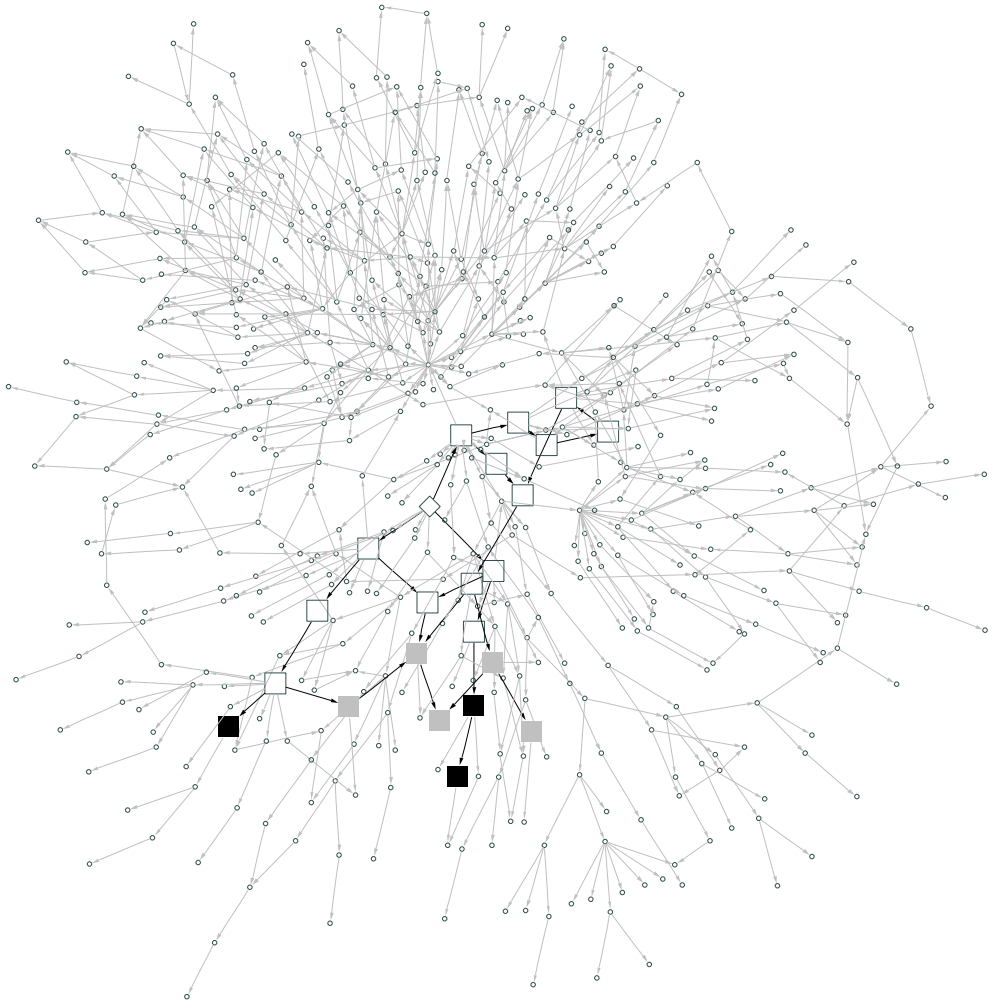


Figure 5.6: Significantly overrepresented Gene Ontology terms associated with binding sites of the transcription factors Mef-2, Myf and TEF. Overrepresented terms are drawn in the context of all 655 terms assigned to the genes with predicted clusters of binding sites. The diamond shows the root node for biological processes; the boxes correspond to those shown in Fig. 5.5.

Chapter 6

Human siRNA Database (HuSiDa)

Summary

Small interfering RNAs (siRNAs), once incorporated into the RNA-induced silencing complex (RISC), mediate the sequence-specific recognition and cleavage of the corresponding target mRNAs. Therefore, siRNAs have become recently a standard tool in functional genomics, allowing efficient regulation (silencing) of specific genes. However, design of siRNAs is laborious and costly, since only a small fraction of randomly chosen siRNA sequences induces an efficient silencing of a target gene. Therefore, in order to improve the design process and algorithms, HuSiDa – an open-access database of published functional human siRNA sequences has been established [Truss et al., 2005] and made available at `http://www.human-siRNA-database.net`. The database provides sequences of siRNAs as well as important technical details of the corresponding gene silencing experiments, including the method of siRNA generation, transfection reagents and procedures, recipient cell lines, and direct links to published references (PubMed).

The database project has been divided in three separate parts. Scanning of the literature for candidate siRNAs and statistical analysis of the collected sequences are the co-authors contributions. Here, the computational implementation of the database and internal data consistency checks are discussed (section 6.2.1). Furthermore, the web interface for searching and accessing the data will be presented (section 6.2.2).

6.1 Introduction

Alteration of gene transcription, investigated in the previous chapters, is the most direct and utilized regulatory tool. Nevertheless, in recent years another cellular mechanism, related rather to gene translation than transcription, attracted efforts of experimental groups. Sequence-specific mRNAs are depleted in a cellular process called RNA interference (RNAi) [Fire et al., 1998]. A ribonuclease III enzyme Dicer cleaves long double-stranded RNA (dsRNA) duplexes or hairpin precursors into fragments of lengths 21–23 nucleotides [Zamore et al., 2000] termed short interfering RNAs (siRNAs). RNA induced silencing complex (RISC) incorporates one of the siRNA strands and cleaves mRNAs containing a subsequence complementary to the attached siRNA strand [Nykanen et al., 2001]. These processes constitute the core of a powerful technique allowing silencing of specific genes to very low levels, and therefore providing a tool for studying gene functions. In most cases, the RNAi effect lasts for 3–5 cell doublings, and normal gene expression resumes in 7–10 cell doublings [McManus and Sharp, 2002].

siRNA molecules for RNAi experiments can be generated by chemical or by enzymatic synthesis. Alternatively, expression vectors can be employed that encode short hairpin RNAs which are intracellularly processed into functional siRNAs by the ribonuclease Dicer. Unfortunately, in mammalian cells these techniques have been shown to induce the interferon response in a concentration-dependent manner, which yields non-specific degradation of mRNAs. Therefore protocols for efficient siRNA delivery into the cell lines of interest need costly optimization.

It has been observed, that only a fraction of siRNA sequences randomly chosen from an mRNA of interest has the capability of inducing an efficient silencing of the corresponding gene. Two siRNAs that target mRNA sites separated by only a few nucleotides may have very different efficacies [Holen et al., 2002, Reynolds et al., 2004]. Moreover, mRNAs containing sequences similar to the exact target sequence may also be influenced. Observations suggest, that central mismatches gradually reduce the silencing effect, whereas at the 5' end of a siRNA the mismatch tolerance is higher [Amarzguioui et al., 2003].

Several attempts have been made to recognize properties of the target sites associated with efficient siRNA silencing. Design rules as well as prediction algorithms have been proposed based on the statistical analysis of the correlation between siRNA sequence and efficacy. These approaches take into account siRNA nucleotide composition [Ui-Tei et al., 2004, Amarzguioui and Prydz, 2004], mRNA secondary structure [Reynolds et al., 2004], duplex stability profiles [Khvorova et al., 2003] and biological information (for

example localization of the translation start sites). A recently published comparative analysis revealed significant differences in the performance of these algorithms [Saetrom and Snove, 2004], demonstrating that the best predictions are obtained by scoring weighted sums of sequence patterns (generated with boosted genetic programming techniques, Saetrom [2004]). It has been pointed out, that for future studies a large independent test/training set of verified siRNA sequences targeting a wide collection of genes is needed.

The availability of functional siRNA sequences and transfection protocols is steadily increasing due to the rapidly growing number of published applications of RNAi. This prompted us [Truss et al., 2005] to establish the database of the siRNA molecules and important technical details of the corresponding gene silencing experiments. In the following the implementation of the database as well as the web interface are presented.

6.2 Results

6.2.1 HuSiDa – database

The HuSiDa database has been organized as a set of tables and a collection of processes for manipulating the table contents. The MySQL system was chosen as the platform to store the data, and the supporting software was written in Perl and SQL. Since the nature of the database requires manual processing of texts of publications, care has been taken to optimize this process.

On a regular basis the database of published articles (PubMed) is searched for publications containing siRNA-related words ("siRNA", "RNAi", "RNA interference") in their titles, abstracts or keywords lists. A list of new candidate articles is prepared for manual processing. Then, an editor of the database decides for each article whether it contains relevant information. If not, the article is marked as rejected and becomes excluded. Otherwise, for each siRNA sequence reported in the article a new siRNA record is created. The editor fills the record fields for the published siRNA sequence, its efficiency, preparation and transfection methods.

Next, each new record is checked. The sequences of biologically non-redundant human mature mRNA sequences (downloaded from the NCBI Reference Sequence (RefSeq) database[1], Pruitt et al. [2005]) are scanned for subsequences matching the manually provided siRNA sequence. Since there is no common standard, stating whether the sense or the antisense siRNA

[1]The file available at `ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz` was used.

sequence should be reported, the direct and the complementary matches to the mRNA sequences are studied. Afterward, all the matches are shown to the editor for checking which one correspond to the gene reported in the original publication. If the list does not contain the target gene, the editor may reject the record or decide to mark it as inconsistent. Otherwise, the record is given a new accession number and added to the HuSiDa database.

6.2.2 HuSiDa – web interface

The information stored in the Human siRNA Database is publicly available at the address `http://www.human-siRNA-database.net`. The contents of the database are presented through dynamically generated hypertext pages (implemented in the PHP4 language). The following main views are provided: a siRNAs searcher, a browser of siRNAs, a viewer of a siRNA record and a viewer of siRNAs listed in a publication. Moreover, a simple user management system allows to provide additional functionality for database maintainers. This way inserting a new siRNA record to the database, editing an existing siRNA record as well as accessing the lists of the publications for manual processing is possible.

In Fig. 6.1 a part of the form for searching siRNAs is presented. The search interface allows to find siRNA records matching a requested gene name, a cell line, a transfection method, a siRNA generation method, or a given author. Alternatively, it is possible to jump to a requested siRNA record based on its accession number. Finally, the searcher page provides a list of all the transfection methods, the siRNA generation techniques and the silenced genes present in the database. After the form is completed and processed, the user is redirected to the siRNAs browser page displaying the results matching the search criteria.

Fig. 6.2 shows an example of search results. They are displayed in a form of a table providing a summary of each siRNA record: a name of the silenced gene, its clickable accession numbers in well-known databases, as well as a PubMed link to the original paper based on which the record was constructed. This table is also used when a user requests to display all the records available in the database. For convenience, the visualized contents of the table can be easily downloaded in a form of a simple text file in the "CSV" format (which is compatibile with majority of popular spreadsheet programs).

Finally, clicking a siRNA record identifier redirects the user to a page containing a detailed description of the selected record. An example of such a page is shown in Fig. 6.3. Here, besides the data described above, the specific siRNA sequence is shown in the form provided in the corresponding

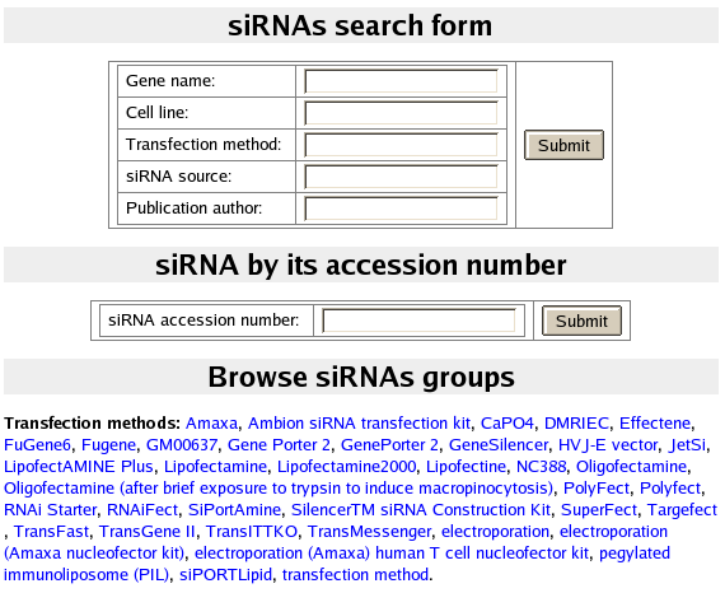


Figure 6.1: Human siRNA Database search form. The records might be selected by a gene name, a cell line, a transfection method, a siRNA generation method or an author. A direct jump to a record with a given accession number is also possible, as well as choosing a gene from a list, etc.

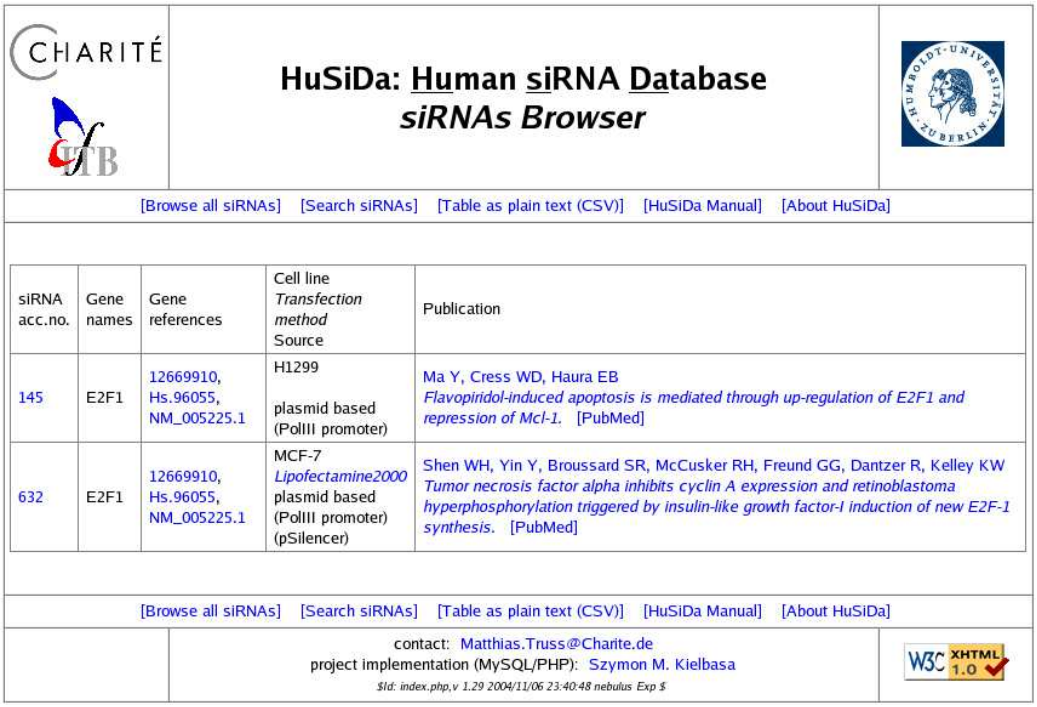
manuscript. Additionally, a list of all mRNA sequences containing the siRNA sequence is given accompanied by the mRNA sequence identifiers and the precise locations of the siRNA matches.

Sometimes, a user might be interested to list all siRNAs described in a given publication. The article browser displays details of all relevant siRNA records in the same form as in Fig. 6.3.

6.3 Discussion

The growing field of the siRNA prediction algorithms requires an independent source of a large set of curated siRNA sequences. At the moment of publication [Truss et al., 2005] the HuSiDa database contained 1158 siRNA records targeting more than 700 different human genes, being the largest collection publicly available (up to the best knowledge of the authors).

It is our intention to facilitate the refinement of the design algorithms by providing our collection of easily accessible sequences. Therefore, the development of the new release of the database takes into account requests of



CHARITÉ

HuSiDa: Human siRNA Database
siRNAs Browser

[Browse all siRNAs] [Search siRNAs] [Table as plain text (CSV)] [HuSiDa Manual] [About HuSiDa]

siRNA acc.no.	Gene names	Gene references	Cell line *Transfection method* Source	Publication
145	E2F1	12669910, Hs.96055, NM_005225.1	H1299 plasmid based (PolIII promoter)	Ma Y, Cress WD, Haura EB *Flavopiridol-induced apoptosis is mediated through up-regulation of E2F1 and repression of Mcl-1.* [PubMed]
632	E2F1	12669910, Hs.96055, NM_005225.1	MCF-7 *Lipofectamine2000* plasmid based (PolIII promoter) (pSilencer)	Shen WH, Yin Y, Broussard SR, McCusker RH, Freund GG, Dantzer R, Kelley KW *Tumor necrosis factor alpha inhibits cyclin A expression and retinoblastoma hyperphosphorylation triggered by insulin-like growth factor-I induction of new E2F-1 synthesis.* [PubMed]

[Browse all siRNAs] [Search siRNAs] [Table as plain text (CSV)] [HuSiDa Manual] [About HuSiDa]

contact: Matthias.Truss@Charite.de
project implementation (MySQL/PHP): Szymon M. Kielbasa
$Id: index.php,v 1.29 2004/11/06 23:40:48 nebulus Exp $

W3C XHTML 1.0

Figure 6.2: A browser of all HuSiDa records or those matching a user query. The texts are linked to the relevant resources: detailed siRNA records, sequence databases and PubMed.

the users. It has been pointed out, that a reliable reference set of siRNA sequences needs to contain the functional records, as well as records experimentally known to induce no (or low) silencing. To address this request, a field describing silencing efficiency has been added to the siRNA record (and missing values are gradually updated). Also a set of forms is under construction, which will allow users to supply their siRNA sequences directly to the database. This approach should speed up the growth of the database. Moreover, it gives a chance to get more sequences of low efficiency, which typically do not belong to the main text of manuscripts.

As it has been presented in the statistical part of Truss et al. [2005], the siRNA sequences stored in the database do not support many of siRNA design criteria proposed before. Soon the number of collected sequences of high efficiency ($> 80\%$ of silencing) should be high enough to provide new and statistically significant insights into a siRNA binding model.

A library, which for each human gene (including potenitially different transcriptional variants) could provide an individual siRNA sequence able to

siRNA record (acc.no. '145')

gene_desc	E2F1 E2F transcription factor 1
gene_name	E2F1
pos	492
strand	+
refs	12669910, Hs.96055, NM_005225.1
sirna_seq_article	GGGGAGAAGTCACGCTATGAGACCTCACTG
sirna_seq_mrna	GGAGAAGTCACGCTATGAG
cell_line	H1299
transfection_method	
sirna_source	plasmid based (PolIII promoter)
efficiency	
sirna_id	145
pmid_id	12533675
mrna_pos_strand_id	E2F1#492#+
insert_user_id	
insert_date	
update_user_id	
update_date	

Figure 6.3: Detailed information available for siRNA records in HuSiDa. The original siRNA sequence is provided in the same form as in the article. All mRNAs of the human genome containing such a sequence are listed.

down-regulate the gene effectively, would be a valuable source for design of experiments. The HuSiDa library has a chance to become such a collection, helpful for studying gene regulation changes caused by turning down single genes.

Chapter 7

Outlook

Gene regulation means temporal and spatial coordination of translation of the genetic material encoded in DNA into proteins. Knowledge of the underlying mechanisms is essential for a deeper understanding of the most fundamental aspects of life like development, differentiation and adaptation to various environmental conditions, the cancer growth. In eukaryotes, gene regulation is predominantly performed at the transcriptional level and it is known, however, that other levels (for example splicing of primary transcripts, RNA stability, RNA depletion directed by siRNAs) are also subject to regulation. A knowledge of position and function of individual regulatory elements as well as their interplay is a prerequisite for the understanding of how cells work in order to eventually form a functioning (or in case of disease: not optimally functioning) organism.

In chapter 2 an algorithm for finding short over-represented words in a given set of promoter sequences was presented. The core idea of the search was to use a degenerated alphabet, which allowed symbols representing multiple nucleotides. This approach appeared to be a good compromise between local alignment methods (working well with many promoter sequences) and methods searching for overrepresented words expressed with the simple `ACGT` alphabet (which could not handle high variability of binding sites). Studies of yeast *Saccharomyces cerevisiae* co-regulated gene families performed with this approach revealed regulatory profiles matching experimentally confirmed ones in 9 out of 11 cases.

A direct application of this technique to a set of co-expressed human genes did not provide satisfactory results. Therefore, different approaches to identify regulatory regions in two human sets of genes were investigated in details, as described in chapter 3. No single method appeared to be capable enough to deliver satisfactory predictions. Nevertheless, integrating methods quantifying independent binding site properties appeared to be a

possible way of gradually improving the quality of the predictions. This observation motivated the development of two novel methods providing further improvements in the predicting algorithms.

Binding site prediction methods typically require a set of profiles recognized by transcription factors at the input. Consequently, dependences in the input set influence the prediction results. Similar profiles lead to predictions of similar binding sites. This increases the risk of an incorrect interpretation of the results if the profile similarity is not a biological property, but rather a result of a profile construction (an alignment method, stringency criteria, etc.). Therefore, to simplify the interpretation, preselection of the profiles is essential. The method presented in the chapter 4 allows to select the representative profiles from a large library. The advantage of this approach lays in a novel combination of two similarity measures, which rely on different profile properties but both take into account the (typically small) number of binding sites used to align the profile. Moreover, both approaches as well as their composition are intuitive and simple, and their computational implementation is straightforward. As an illustration of the method a comparison of the Jaspar and Transfac libraries identifies nearly identical profiles, which are difficult to find without expert knowledge. Studies of cooperative binding of multiple transcription factors provide a potential future application of the method, since they require to eliminate artificial binding site pairs caused by too similar profiles.

In order to reduce the number of false signals, tools predicting binding sites in higher eukaryotic sequences quantify associations of candidate sites with other predicted sites, gene expression levels, evolutionary conserved sequences, etc. The TFGossip algorithm, presented in chapter 5, proposes to include a next category of information – gene functional annotations originating from the Gene Ontology project. It is demonstrated, that this novel technique is able to predict correctly biological processes regulated by a few well-known sets of transcription factors, keeping the rate of falsely discovered processes below a given threshold. Functions of a single transcription factor and cooperating transcription factors can be studied. Moreover, as it is mentioned, filtering predicted binding sites by observing whether the corresponding genes are annotated with the discovered terms improves the site predictions. Therefore, a natural next step is to incorporate a TFGossip-like approach into a larger binding site prediction framework. Such a framework should combine predictions of binding sites coming from different algorithms, taking into account that some of the signals may be mutually dependent and that the rate of false positivies is high. Ideally, this approach should be optimized for studies of gene sets of moderate size (10-100) and should provide a possibility to incorporate manually entered expert knowledge. The amount

of available expermiental data, which describes gene co-expression, should be already sufficient to train such a complex probabilistic framework reliably.

Furthermore, TFGossip might be the first step toward understanding of the functions regulated by combinatorial interactions of known transcription factors on a genome-wide scale. A map could be constructed assigning to pairs (or larger sets) of transcription factors biological processes regulated by them. Here, the technique described above, for choosing a core subset of profiles recognized by transcription factors could be used to reduce the number of factor pairs. Of course, in this context multiple testing issue would require further studies.

Finally, the majority of the methods refereed in this thesis analyse sets of co-regulated genes. In recent years, experiments employing RNA interference have become a source of lists containing genes up- or down-regulated as an effect of a knock-down of a specific target. The difficulties in prediction, which position of mRNA should be targeted by a short interfering RNAs in order to achieve an efficient down-regulation of the corresponding gene, prompted the development of the curated database of active human siRNA sequences (and their transfection methods). The database, presented in details in chapter 6, at the moment of publication contained the largest publicly available set of active sequences. The database might become a reference set for comparing effects on gene regulation of siRNAs predicted by newly designed algorithms.

Further extensions of the approaches discussed in the chapters of the thesis might result in algorithms powerful enough to provide reliable predictions of a quality high enough to be verified experimentally. Incorporation of such predictions into studies of small regulatory networks (for example control of the cell cycle or circadian clock) might help to identify potential interactions. Additionally, in studies of signal transduction (like the RAS cascade) predicted binding sites might help in distinguishing first target genes from those activated by the next steps of the cascade. In order to achieve this goal, rigorous quality control steps have to be implemented in the prediction pipelines. Introduction of such controls requires a strict cooperation between computational specialists and experimental biologists to guarantee statistical correctness and biological relevance of studied values. Last but not least, providing intuitive visualization techniques would be a further achievement in the field.

Appendix A

Overrepresented words as regulatory elements

A.1 Z-score formula

In section 2.2.1, as the measure of biological importance of a motif W the Z-score formula (Eq. 2.1) is used (W represents here a short motif expressed with the the alphabet containing the symbols for nucleotides `ACGT` as well as the symbols representing their mixtures). A modified version of the approach introduced first by Pevzner et al. [1989] is used here to calculate $\sigma(W)$. For the `ACGT` alphabet the correction term for self-overlapping words has no large influence on the ranked list of motifs. However, when the degenerated alphabet is used, the number of motif self-overlaps increases and both terms in $\sigma(W)$ become comparable possibly influencing the order on the top motifs list.

In order to find the final formula for $Z(W)$, a set of patterns $\mathcal{W}$ equivalent to a certain motif W (of length L_W bases) needs to be defined. The set includes all `ACGT`-like patterns which match the motif W or the motif complementary to it. For example, if $W =$ `CAWTCA` then $\mathcal{W}$ contains `TGATTG`, `CAATCA`, `TGAATG`, `CATTCA`.

Let's assume that the motif occurrences are searched in a family of M promoter sequences, each of the length N_i+L_W-1 nucleotides ($i = 1, \ldots, M$), so the total number of positions where a match is possible equals to: $N = \sum_{i=1}^{M} N_i$. Note, that if the length L_W of motifs is increased by k bases, the number of positions decreases by $M \cdot k$.

The number of observations $n_{\mathrm{obs}}(W)$ is defined as the total number of positions in the promoters, where a pattern belonging to the set $\mathcal{W}$ is observed.

The expected values $\mu(W)$ and $\sigma(W)$ are calculated based on a large set of promoters, independent of the family. This set in conjunction with a background model trained on it is used to calculate reference probabilities $p(w)$ of any ACGT-type pattern w of a length in range from L_W to $2L_W - 1$ bases. Here, as the background model Markov models of different order are used.

The mean

The expression $w \to i$ means that the pattern w is present at the position i. Let's define x_i^w to be a random variable equal to:

$$x_i^w = \begin{cases} 1, & w \to i \\ 0, & w \not\to i \end{cases} .$$

Next, a random variable describing the total number of occurrences of patterns w matching the motif W anywhere in the co-regulated family is defined as follows:

$$X(\mathcal{W}) = \sum_{w \in \mathcal{W}} \sum_{i=1}^{N} x_i^w .$$

The average over all ensembles $< X(\mathcal{W}) >$ is equal to the expected number of occurrences of the motif $\mu(W)$. Assuming stationarity we can calculate the averages using the corresponding word probabilities $< x_i^w > = p(w)$ and then:

$$\mu(W) = < X(\mathcal{W}) > = \sum_{w \in \mathcal{W}} N p(w).$$

The variance

Applying the definition of the variance we have:

$$\begin{aligned} \sigma^2(X) &= < X^2 > - < X >^2 \\ &= \sum_{i,j=1}^{N} \sum_{w,v \in \mathcal{W}} \Big(< x_i^w x_j^v > \\ &\quad - < x_i^w >< x_j^v > \Big) . \end{aligned}$$

Using the symmetry of this formula under exchange of i and j the double sum over positions may be rewritten as:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} f_{i,j} = \sum_{i=1}^{N} f_{i,i} + 2 \sum_{i=1}^{N} \sum_{s=1}^{N-i} f_{i,i+s},$$

using the abbreviation:

$$f_{i,j} = \sum_{w,v \in \mathcal{W}} \left(< x_i^w x_j^v > - < x_i^w >< x_j^v > \right).$$

Note that it is impossible to have two different motifs in the same position, so the term $f_{i,i}$ can be reduced to:

$$f_{i,i} = \sum_{w \in \mathcal{W}} \left(< x_i^w x_i^w > - < x_i^w >^2 \right).$$

The possible values of the random variable $x_i^w x_i^w$ are:

$$x_i^w x_i^w = \begin{cases} 1, & w \to i \\ 0, & w \not\to i \end{cases},$$

so after averaging over all ensembles $< x_i^w x_i^w > = p(w)$ we get the binomial part of the expression for the standard deviation:

$$\sum_{i=1}^{N} f_{i,i} = \sum_{w \in \mathcal{W}} Np(w) \left(1 - p(w)\right).$$

In order to calculate the term of the variance containing $f_{i,i+s}$, the meaning of s (the offset) should be clarified. If $s \geq L_W$ then there is no overlap between the patterns located at the positions i and $i+s$ and we assume that we can treat these patterns as uncorrelated (generalizations for correlated sequences can by found in Kleffe and Borodovsky [1991]):

$$f_{i,i+s} = 0 \quad \text{for} \quad s \geq L_W.$$

In the remaining double sum over positions $2 \sum_{i=1}^{N} \sum_{s=1}^{N-i} f_{i,i+s}$ it is enough to take into account the range of s from 1 to $L_W - 1$ when words overlap. The random variable

$$x_i^w x_{i+s}^v = \begin{cases} 1, & w \to i \wedge v \to i+s \\ 0, & \text{otherwise} \end{cases}$$

equals one only when in the patterns w and v the letters on overlapping positions are the same giving a word $Q_s^{w,v}$ of the length $L_W + s$. Then averaging

$$< x_i^w x_{i+s}^v > = p(Q_s^{w,v}) \equiv \pi_s^{w,v}.$$

If the letters disagree (no overlap), $\pi_s^{w,v} = 0$.

Here we note, that there are N positions for motifs of length L_W. Increasing the word length to $L_W + s$ reduces the number of possible positions by $s \cdot M$. Consequently we may write:

$$2 \sum_{i=1}^{N} \sum_{s=1}^{N-i} f_{i,i+s} =$$

$$2 \sum_{s=1}^{L_W - 1} (N - sM) \sum_{w,v \in \mathcal{W}} (\pi_s^{w,v} - p(w)p(v))$$

and finally we obtain the formula (Eq. 2.2) for the variance of the motif W.

Appendix B

Lists of similar profiles

As presented in chapter 3 a careful selection of independent transcription factor profiles is needed. A statistical approach eliminating a potential bias caused by a manual profile selection is discussed in chapter 4. Using this method the profiles available in well known libraries Jaspar [Sandelin et al., 2004a] and Transfac [Wingender et al., 1996, 2000, Matys et al., 2003] were compared. In Tab. B.1 the matrices provided by the smaller Jaspar library are mapped to the records of the Transfac library. The second Tab. B.2 provides a list of clusters of similar matrices within the set containing the matrices of both databases.

Jaspar	Transfac
V_HNF-1	V$HNF1_01, V$HNF1_Q6
V_NRF-2	V$ELK1_02, V$NRF2_01
V_c-ETS	V$PEA3_Q6
V_RREB-1	V$RREB1_01
P_Dof2	P$DOF1_01, P$DOF2_01, P$DOF3_01, P$PBF_01, V$PAX2_02
V_Hen-1	V$HEN1_02, V$HEN1_01, V$LBP1_Q6
V_HFH-2	V$FOXD3_01
V_E4BP4	V$CREBP1_01, V$E4BP4_01
X_NF-Y	VNFY_01, VALPHACP1_01, V$ALPHACP1_01
P_bZIP910	V$CREB_01, V$CREBP1CJUN_01, P$BZIP910_02, V$ATF6_01
V_SPI-B	V$PU1_Q6
V_p50	V$NFKAPPAB50_01
V_Pax6	V$PAX6_01
P_bZIP911	P$BZIP911_01, V$ATF6_01
V_SRF	V$SRF_01
V_SPI-1	V$PU1_Q6
V_Irf-2	V$IRF1_01, V$IRF2_01
V_p65	V$NFKAPPAB65_01, V$CREL_01, V$NFKB_Q6
V_Sox-5	V$SOX5_01, V$SRY_02, V$SOX9_B1
I_Dorsal_2	IDL_01, VNFKAPPAB65_01, V$CREL_01
V_Pbx	V$PBX1_02
V_n-MYC	V$NMYC_01, F$PHO4_01, V$MYCMAX_01, V$MAX_01, VUSF_01, VUSF_02, V$MYCMAX_02, V$SREBP1_01, V$ARNT_01, P$EMBP1_Q2, P$CPRF_Q2, $ARNT_02, V$MYCMAX_03, V$MYC_Q2
V_Irf-1	V$IRF1_01, V$IRF2_01, V$ICSBP_Q6
V_SP1	V$SP1_01
V_Max	V$NMYC_01, V$MAX_01, VUSF_01, VUSF_02, V$MYCMAX_02, V$ARNT_01, P$EMBP1_Q2, P$HBP1A_Q2, P$TAF1_Q2, P$CPRF2_Q2, V$MYCMAX_03, V$USF_Q6_01, VMYC_Q2, PCPRF2_01

Table B.1: Correspondence between Jaspar and Transfac matrices. For each Jaspar matrix similar Transfac matrices ($D \leq 1$ and $C \geq 0.8$) are listed. 84 Jaspar matrices have at least one corresponding Transfac matrix.

Jaspar	Transfac
V_SOX-9	V$SOX9_B1
V_HFH-1	V$HFH1_01
V_USF	V$NMYC_01, F$PHO4_01, V$MYCMAX_01, V$MAX_01, VUSF_01, VMYCMAX_02, V$SREBP1_01, V$ARNT_01, P$EMBP1_Q2, P$CPRF_Q2, V$ARNT_02, V$MYCMAX_03, VMYC_Q2, PTAF1_01
P_Dof3	P$DOF2_01, P$DOF3_01, P$PBF_01
V_CREB	V$CREB_01, V$CREB_02, V$CREB_Q4_01
V_AML-1	V$AML1_01, V$COREBINDINGFACTOR_Q6, V$AML1_Q6, V$AML_Q6
P_AGL3	P$AGL3_01, P$AGL3_02
I_CFI-USP	I$CF1_01, I$CF1_02
V_AP2alpha	V$AP2ALPHA_01
V_FREAC-2	V$FREAC2_01
V_PPARgamma	V$PPARG_02
V_p53	V$P53_01
P_Agamous	P$AG_01
I_Broad-complex_1	I$BRCZ1_01
V_deltaEF1	V$DELTAEF1_01, V$AREB6_02
V_Staf	V$STAF_02
V_Nkx	V$NKX25_01, V$NKX25_02
V_MEF2	V$RSRFC4_01
V_GATA-1	V$GATA1_01, V$GATA2_01
V_Ahr-ARNT	V$AHRARNT_01, V$AHR_Q5
I_Broad-complex_2	I$BRCZ2_01
V_MZF_1-4	V$MZF1_01
I_E74A	I$E74A_01, V$ELK1_02, V$CETS1P54_01, V$CETS1P54_02, V$NRF2_01, V$CETS168_Q6
V_RORalfa-1	V$RORA1_01, V$ERR1_Q2, V$ER_Q6_02
P_Athb-1	P$ATHB1_01, P$ATHB5_01
I_CF2-II	I$CF2II_01, I$CF2II_02
V_Elk-1	V$ELK1_02, V$CETS1P54_01, V$NRF2_01
V_HLF	VVBP_01, VHLF_01
X_TBP	V$TATA_01
V_SRY	V$SOX5_01, V$SRY_02
V_Myc-Max	V$MYCMAX_01, V$MYC_Q2

Jaspar	Transfac
V_NF-kappaB	V$NFKAPPAB65_01, V$NFKAPPAB_01, V$NFKB_Q6
V_c-REL	V$NFKAPPAB65_01, V$CREL_01
V_COUP-TF	V$COUP_01, V$HNF4ALPHA_Q6, V$PPAR_DR1_Q2, V$HNF4_DR1_Q3, V$COUP_DR1_Q6
I_Snail	ISN_01, VE47_02, V$LMO2COM_01, V$E12_Q6, V$MYOD_Q6_01
I_Broad-complex_3	I$BRCZ3_01
V_Gfi	V$GFI1_01
V_GATA-3	V$GATA3_01, V$GATA1_02, V$GATA6_01
V_Chop-cEBP	V$CHOP_01
V_ARNT	V$NMYC_01, F$PHO4_01, V$MYCMAX_01, V$MAX_01, VUSF_01, VMYCMAX_02, V$ARNT_01, P$EMBP1_Q2, P$CPRF_Q2, P$TAF1_Q2, V$ARNT_02, V$MYCMAX_03, VMYC_Q2, PTAF1_01
V_HNF-3beta	V$HNF3B_01, V$HNF3ALPHA_Q6
I_SU_h	I$SUH_01
I_Dorsal_1	I$DL_01
V_PPARgamma-RXRal	V$PPARG_01
V_FREAC-4	V$XFD3_01, V$FREAC2_01
V_Thing1-E47	V$HAND1E47_01
V_RORalfa-2	V$RORA2_01
V_Yin-Yang	V$YY1_Q6
I_Hunchback	I$HB_01
V_c-MYB_1	V$CMYB_01
V_TCF11-MafG	V$AP1FJ_Q2, V$AP1_Q6
P_GAMYB	P$GAMYB_01
V_SAP-1	V$CETS1P54_01, V$NRF2_01, V$CETS168_Q6
V_Bsap	V$PAX5_01
V_HFH-3	V$HFH3_01
V_MZF_5-13	V$MZF1_02
V_Evi-1	V$EVI1_06, V$EVI1_01, V$EVI1_02, V$EVI1_03, V$EVI1_05
V_E2F	V$E2F_02, V$E2F_Q3, V$E2F_Q4, V$E2F_Q6, V$E2F1_Q3, V$E2F1_Q4, V$E2F1_Q6, V$E2F_03, V$E2F1DP1_01, V$E2F1DP2_01, V$E2F4DP1_01, V$E2F4DP2_01, V$E2F1DP1RB_01, V$E2F_Q3_01, V$E2F1_Q4_01, V$E2F1_Q6_01
V_c-FOS	V$AP1_Q6, V$AP1_Q4, V$BACH2_01, V$AP1_01

Cluster	Matrices
1	I$CF2II_02, I$CF2II_01, I_CF2-II
2	V$CETS1P54_01, V$ELK1_02, V$CETS1P54_02, V$NRF2_01, V$CETS168_Q6, V_NRF-2, I_E74A, I$E74A_01, V_Elk-1, V_SAP-1
3	V$CREB_01, V$ATF_01, V$CREBP1CJUN_01, V$VJUN_01, V$CREB_Q2, V$CREB_Q4, V$CREBP1_Q2, V$CREB_02, P$BZIP911_01, P$BZIP910_01, V$ATF6_01, V$ATF3_Q6, V$ATF4_Q2, V$CREB_Q4_01, P_bZIP910, P$BZIP910_02, P_bZIP911, V_CREB
4	V$E4BP4_01, V$CREBP1_01, V_E4BP4
5	V$CREL_01, V$NFKAPPAB65_01, V$NFKAPPAB_01, V$NFKB_Q6, V_p65, I_Dorsal_2, I$DL_01, V_NF-kappaB, V_c-REL, I_Dorsal_1
6	V$IRF2_01, V$IRF1_01, V_Irf-2, V_Irf-1, V$ICSBP_Q6
7	V$TAL1ALPHAE47_01, V$TAL1BETAE47_01, V$TAL1BETAITF2_01
8	V$HEN1_01, V$HEN1_02, V$AP4_Q6, V$AP4_Q5, V$LBP1_Q6, V_Hen-1
9	V$GATA2_01, V$GATA1_01, V_GATA-1
10	V$EVI1_02, V$EVI1_06, V$EVI1_03, V$EVI1_01, V$EVI1_05, V_Evi-1
11	V$CLOX_01, V$CDP_02
12	V$CDPCR3HD_01, V$CDPCR1_01
13	I$CF1_02, I$CF1_01, I_CFI-USP
14	V$CEBPB_02, V$CEBPA_01, V$CEBP_Q2, V$CEBP_Q2_01
15	VUSF_01, VMAX_01, V$ARNT_01, P$EMBP1_Q2, PGBP_Q6, PHBP1A_Q2, VUSF_Q6, PCPRF_Q2, V$NMYC_01, P$TAF1_Q2, P$CPRF3_Q2, P$CPRF2_Q2, P$O2_02, P$TGA1B_Q2, P$TGA1A_Q2, P$GBF_Q2, PABF_Q2, PABF1_01, P$O2_Q2, V$MYCMAX_03, P$RITA1_01, V$E4F1_Q6, P$HBP1B_Q6, V$USF2_Q6, VUSF_02, VMYC_Q2, F$PHO4_01, V$MYCMAX_01, P$HBPA1_Q6_01, P$ROM_Q2, P$TAF1_01, P$CPRF3_01, P$CPRF2_01, V_n-MYC, V$MYCMAX_02, V$SREBP1_01, V$ARNT_02, V_Max, V$USF_Q6_01, V_USF, V_Myc-Max, V_ARNT
16	V$GATA1_02, V$GATA3_01, V$GATA1_04, V$GATA1_03, V$LMO2COM_02, V$GATA1_05, V$GATA1_06, V$GATA2_02, V$GATA2_03, V$GATA3_02, V$GATA6_01, V$GATA_Q6, V_GATA-3

Table B.2: Clusters of similar ($D \leq 1$ and $C \geq 0.8$) Jaspar and Transfac matrices.

Cluster	Matrices
17	V$HNF3B_01, V$FOXD3_01, V$HFH3_01, V$HNF3ALPHA_Q6, V$FOX_Q2, V_HFH-2, V_HNF-3beta, V_HFH-3
18	IBCD_01, IDFD_01
19	V$HSF2_01, V$HSF1_01
20	VSRY_02, VSOX5_01, V_Sox-5, V$SOX9_B1, V_SOX-9, V_SRY
21	V$AP1_Q2, V$AP1FJ_Q2, V$AP1_Q6, V$AP1_Q4, V$BACH1_01, V$BACH2_01, V$AP1_01, V$AP1_Q2_01, V$AP1_Q6_01, V$AP1_Q4_01, V_TCF11-MafG, V_c-FOS
22	V$E2_Q6, V$E2_01
23	VHLF_01, VVBP_01, V_HLF
24	V$XFD2_01, I$CROC_01, V$FREAC2_01, V$FREAC7_01, V_FREAC-2, V_FREAC-4, V$XFD3_01
25	V$LMO2COM_01, V$MYOD_01, V$E12_Q6, V$E47_01, ISN_01, VMYOGENIN_Q6, VHEB_Q6, VAP4_Q6_01, V$MYOD_Q6_01, I_Snail, V$E47_02
26	V$TCF11_01, N$SKN1_01
27	V$HFH8_01, V$HFH1_01, V_HFH-1
28	VGEN_INI3_B, VGEN_INI2_B, V$GEN_INI_B
29	V$MINI19_B, V$MUSCLE_INI_B, V$MINI20_B
30	P$DOF2_01, P$DOF1_01, P$DOF3_01, P$PBF_01, V$PAX2_02, P_Dof2, P_Dof3
31	V$MMEF2_Q6, V$AMEF2_Q6, V$HMEF2_Q6, V$MEF2_01
32	V$HNF4_01_B, V$HNF4_01, V$PPARG_03, V$PPARA_01, V$DR1_Q3
33	V$AREB6_02, V$DELTAEF1_01, V_deltaEF1
34	V$MEIS1BHOXA9_01, V$MEIS1AHOXA9_01
35	V$E2F_Q3, V$E2F_02, V$E2F_Q4, V$E2F_Q6, V$E2F1_Q3, V$E2F1_Q4, V$E2F1_Q6, V$E2F_03, V$E2F1DP1_01, V$E2F1DP2_01, V$E2F4DP1_01, V$E2F4DP2_01, V$E2F1DP1RB_01, V$E2F_Q2, V$E2F_Q3_01, V$E2F1_Q4_01, V$E2F1_Q6_01, V_E2F
36	V$ZIC2_01, V$ZIC1_01, V$ZIC3_01
37	V$STAT5B_01, V$STAT5A_01
38	V$POU3F2_02, V$OCT1_07
39	V$FOXO1_01, V$FOXO4_01
40	N$DAF16_01, V$FOXO1_02, V$FOXO4_02, V$FOXO3_01
41	V$NKX22_01, V$NKX25_01, V_Nkx, V$NKX25_02

Cluster	Matrices
42	V$STAT6_01, V$STAT5A_03, V$STAT1_03, V$STAT3_02, V$STAT4_01, V$STAT5A_04, V$STAT6_02
43	P$ATHB5_01, P$ATHB1_01, P_Athb-1
44	V$ERR1_Q2, V$RORA1_01, VER_Q6_02, VT3R_01, V_RORalfa-1
45	V$HNF4ALPHA_Q6, V$COUP_01, V$PPAR_DR1_Q2, V$HNF4_DR1_Q3, V$COUP_DR1_Q6, V_COUP-TF
46	V$ALPHACP1_01, V$NFY_01, VNFY_Q6_01, VNFY_Q6, X_NF-Y
47	V$PAX8_01, V$PAX8_B
48	V$COREBINDINGFACTOR_Q6, V$AML1_01, V$AML1_Q6, V$AML_Q6, V_AML-1
49	F$ROX1_Q6, F$MAT1MC_02
50	V$P53_DECAMER_Q2, V$P53_02
51	VETS_Q4, VPEA3_Q6, V$TEL2_Q6, V_c-ETS
52	VIRF_Q6, VISRE_01
53	V$SREBP_Q3, V$SREBP1_02, V$SREBP1_Q6
54	V$HNF1_Q6, V$HNF1_01, V_HNF-1
55	VOCT_Q6, VOCT1_05, V$OCT1_Q5_01
56	V$HIF1_Q3, V$HIF1_Q5
57	P$MYBAS1_01, V$MYB_Q6, VMYB_Q5_01, PC1_Q2
58	V$E2A_Q2, V$MYOD_Q6
59	VSRF_Q4, VSRF_Q6, V$SRF_Q5_01
60	V$CREB_Q2_01, V$CREB_Q3
61	V$E2F_Q6_01, V$E2F_Q4_01
62	VGR_Q6_01, VGR_Q6
63	V$SP1_Q6_01, V$SP1_Q6, VETF_Q6, VSP1_Q4_01
64	I$ZESTE_Q2_01, I$ZESTE_Q2
65	V$NFAT_Q4_01, V$NFAT_Q6
66	VPR_02, VGR_01
67	V$DR4_Q2, V$LXR_DR4_Q3
68	V_RREB-1, V$RREB1_01
69	V_SPI-B, V$PU1_Q6, V_SPI-1
70	V_p50, V$NFKAPPAB50_01
71	V_Pax6, V$PAX6_01
72	V_SRF, V$SRF_01
73	V_Pbx, V$PBX1_02

Cluster	Matrices
74	V_SP1, V$SP1_01
75	P_AGL3, P$AGL3_01, P$AGL3_02, P_SQUA
76	V_AP2alpha, V$AP2ALPHA_01
77	V_PPARgamma, V$PPARG_02
78	V_p53, V$P53_01
79	P_Agamous, P$AG_01
80	I_Broad-complex_1, I$BRCZ1_01
81	V_Staf, V$STAF_02
82	V_MEF2, V$RSRFC4_01
83	V_Ahr-ARNT, V$AHRARNT_01, V$AHR_Q5
84	I_Broad-complex_2, I$BRCZ2_01
85	V_MZF_1-4, V$MZF1_01
86	X_TBP, V$TATA_01
87	I_Broad-complex_3, I$BRCZ3_01
88	V_Gfi, V$GFI1_01
89	V_Chop-cEBP, V$CHOP_01
90	I_SU_h, I$SUH_01
91	V_PPARgamma-RXRal, V$PPARG_01
92	V_Thing1-E47, V$HAND1E47_01
93	V_RORalfa-2, V$RORA2_01
94	V_Yin-Yang, V$YY1_Q6
95	I_Hunchback, I$HB_01
96	V_c-MYB_1, V$CMYB_01
97	P_GAMYB, P$GAMYB_01
98	V_Bsap, V$PAX5_01
99	V_MZF_5-13, V$MZF1_02

Bibliography

S. F. Altschul, T. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

M. Amarzguioui and H. Prydz. An algorithm for selection of functional siRNA sequences. *Biochem Biophys Res Commun*, 316(4):1050–8, 2004.

M. Amarzguioui, T. Holen, E. Babaie, and H. Prydz. Tolerance for mutations and chemical modifications in a siRNA. *Nucleic Acids Res*, 31(2):589–95, 2003.

M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximisation to discover motifs in biopolymers. In *Proc. int. conf. intell. syst. mol. biol.*, pages 28–36, Menlo Park, California, 1994. AAAI Press.

O. Berg and P. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–50, 1987.

E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey,

A. Ureta-Vidal, C. Woodwark, M. Clamp, and T. Hubbard. Ensembl 2004. *Nucleic Acids Res*, 32 Database issue:D468–70, 2004.

N. Blüthgen, S. Kielbasa, B. Cajavec, and H. Herzel. HOMGL-comparing genelists across species and with different accession numbers. *Bioinformatics*, 20(1):125–6, 2004.

N. Blüthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, and D. Beule. Biological profiling utilizing gene ontology. *submitted*, 2005a.

N. Blüthgen, S. Kielbasa, and H. Herzel. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res*, 33(1):272–9, 2005b.

A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8:1202–1215, 1998.

H. Bussemaker, H. Li, and E. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001.

L. Cardone and P. Sassone-Corsi. Timing the cell cycle. *Nat Cell Biol*, 5(10): 859–61, 2003.

M. Caselle, F. Di Cunto, and P. Provero. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics*, 3(1):7, 2002.

S. Cawley, S. Bekiranov, H. Ng, P. Kapranov, E. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116(4):499–509, 2004.

W. Chen and R. Baler. The rat arylalkylamine N-acetyltransferase E-box: differential use in a master vs. a slave oscillator. *Brain Res Mol Brain Res*, 81(1-2):43–50, 2000.

H. A. Coller, C. Grandori, P. Tamayo, T. Colbert, E. S. Lander, R. N. Eisenman, and T. R. Golub. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. USA*, 97:3260–3265, 2000.

T. Darlington, K. Wager-Smith, M. Ceriani, D. Staknis, N. Gekakis, T. Steeves, C. Weitz, J. Takahashi, and S. Kay. Closing the circadian

loop: CLOCK-induced transcription of its own inhibitors per and tim. *Science*, 280(5369):1599–603, 1998.

R. Davuluri, I. Grosse, and M. Zhang. Computational identification of promoters and first exons in the human genome. *Nat Genet*, 29(4):412–7, 2001.

L. Diatchenko, Y. Lau, A. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E. Sverdlov, and P. Siebert. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A*, 93(12):6025–30, 1996.

C. Dieterich, B. Cusack, H. Wang, K. Rateitschak, A. Krause, and M. Vingron. Annotating regulatory DNA based on man-mouse genomic comparison. *Bioinformatics*, 18 Suppl 2:S84–90, 2002.

C. Dieterich, H. Wang, K. Rateitschak, H. Luz, and M. Vingron. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*, 31 (1):55–7, 2003.

S. Elbashir, J. Harborth, K. Weber, and T. Tuschl. Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, 26(2): 199–213, 2002.

G. Euskirchen, T. Royce, P. Bertone, R. Martone, J. Rinn, F. Nelson, F. Sayward, N. Luscombe, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol*, 24 (9):3804–14, 2004.

S. Fessele, S. Boehlk, A. Mojaat, N. Miyamoto, T. Werner, E. Nelson, D. Schlondorff, and P. Nelson. Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J*, 15(3):577–9, 2001.

S. Fessele, H. Maier, C. Zischek, P. Nelson, and T. Werner. Regulatory context is a crucial part of gene function. *Trends Genet*, 18(2):60–3, 2002.

J. Fickett. Quantitative discrimination of MEF2 sites. *Mol Cell Biol*, 16(1): 437–41, 1996.

E. Filipski, V. King, X. Li, T. Granda, M. Mormont, X. Liu, B. Claustrat, M. Hastings, and F. Levi. Host circadian clock as a control point in tumor progression. *J Natl Cancer Inst*, 94(9):690–7, 2002.

A. Fire, S. Xu, M. Montgomery, S. Kostas, S. Driver, and C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–11, 1998.

M. Frith, J. Spouge, U. Hansen, and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30(14):3214–24, 2002.

M. Frith, M. Li, and Z. Weng. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*, 31(13):3666–8, 2003.

M. Frith, Y. Fu, L. Yu, J. Chen, U. Hansen, and Z. Weng. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res*, 32(4):1372–81, 2004a.

M. Frith, U. Hansen, J. Spouge, and Z. Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200, 2004b.

N. Gekakis, D. Staknis, H. Nguyen, F. Davis, L. Wilsbacher, D. King, J. Takahashi, and C. Weitz. Role of the CLOCK protein in the mammalian circadian mechanism. *Science*, 280(5369):1564–9, 1998.

B. Gottgens, L. M. Barton, J. G. Gilbert, A. J. Bench, M. J. Sanchez, S. Bahn, S. Mistry, D. Grafham, A. McMurray, M. Vaudin, E. Amaya, D. R. Bentley, A. R. Green, and A. M. Sinclair. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol*, 18(2):181–186, Feb 2000.

S. Hannenhalli and S. Levy. Predicting transcription factor synergism. *Nucleic Acids Res*, 30(19):4278–84, 2002.

P. Haverty, M. Frith, and Z. Weng. CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res*, 32(Web Server issue):W213–6, 2004.

T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A. Kel, O. Kel, E. Ignatieva, E. Ananko, O. Podkolodnaya, F. Kolpakov, N. Podkolodny, and N. Kolchanov. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res*, 26(1):362–7, 1998.

G. Hertz and G. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15 (7-8):563–77, 1999.

S. Herwig and M. Strauss. The retinoblastoma protein: a master regulator of cell cycle, differentiation and apoptosis. *Eur J Biochem*, 246(3):581–601, 1997.

H. Herzel, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, and J. Schuchhardt. Extracting information from cDNA arrays. *Chaos*, 11(1):98–107, 2001.

J. Hogenesch, Y. Gu, S. Jain, and C. Bradfield. The basic-helix-loop-helix-PAS orphan MOP3 forms transcriptionally active complexes with circadian and hypoxia factors. *Proc Natl Acad Sci U S A*, 95(10):5474–9, 1998.

T. Holen, M. Amarzguioui, M. Wiiger, E. Babaie, and H. Prydz. Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res*, 30(8):1757–66, 2002.

Z. Hu, Y. Fu, A. Halees, S. Kielbasa, and Z. Weng. SeqVISTA: a new module of integrated computational tools for studying transcriptional regulation. *Nucleic Acids Res*, 32(Web Server issue):W235–41, 2004.

J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–14, 2000.

A. Kel, O. Kel-Margoulis, V. Babenko, and E. Wingender. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol*, 288(3):353–76, 1999.

A. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–9, 2003.

O. Kel-Margoulis, A. Romashchenko, N. Kolchanov, E. Wingender, and A. Kel. COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.*, 28:311–15, 2000.

A. Khvorova, A. Reynolds, and S. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–16, 2003.

S. Kielbasa, J. Korbel, D. Beule, J. Schuchhardt, and H. Herzel. Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*, 17(11):1019–26, 2001.

S. Kielbasa, N. Blüthgen, and H. Herzel. Genome-wide Analysis of Functions Regulated by Sets of Transcription Factors. *Proceedings of the German Conference on Bioinformatics*, pages 105–113, 2004a.

S. Kielbasa, N. Blüthgen, C. Sers, R. Schäfer, and H. Herzel. Prediction of Cis-Regulatory Elements of Coregulated Genes. *Genome Informatics*, 15 (1):117–124, 2004b.

S. Kielbasa, D. Gonze, and H. Herzel. Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6:237, 2005.

J. Kleffe and M. Borodovsky. First and second moment of counts of words in random texts generated by Markov chains. *CABIOS*, 8, 1991.

L. Kuras, H. Chérest, Y. Surdin-Kerjan, and D. Thomas. A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, met4 and met28, mediates the transcription activation of yeast sulfur metabolism. *EMBO J.*, 15:2519–2529, 1996.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

R. Liu and D. States. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res*, 12(3):462–9, 2002.

R. Martone, G. Euskirchen, P. Bertone, S. Hartman, T. Royce, N. Luscombe, J. Rinn, F. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A*, 100(21):12247–52, 2003.

T. Matsuo, S. Yamaguchi, S. Mitsui, A. Emi, F. Shimoda, and H. Okamura. Control mechanism of the circadian clock for timing of cell division in vivo. *Science*, 302(5643):255–9, 2003.

V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. Kel, O. Kel-Margoulis, D. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–8, 2003.

M. McManus and P. Sharp. Gene silencing in mammals by small interfering RNAs. *Nat Rev Genet*, 3(10):737–47, 2002.

H. W. Mewes, A. K., K. Heumann, S. Liebl, and F. Pfeiffer. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, 25:28–30, 1997.

G. Miller and H. Quastler. *Information theory in psychology.* Free Press, Glencoe, Ill., 1955.

E. Munoz, M. Brewer, and R. Baler. Circadian Transcription. Thinking outside the E-Box. *J Biol Chem*, 277(39):36009–17, 2002.

K. Murakami, T. Kojima, and Y. Sakaki. Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. *BMC Genomics*, 5(1):16, 2004.

A. Nykanen, B. Haley, and P. Zamore. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell*, 107(3):309–21, 2001.

U. Ohler, H. Niemann, Liao Gc, and G. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17 Suppl 1:S199–206, 2001.

R. C. Perier, T. Junier, and P. Bucher. The eukaryotic promoter database EPD. *Nucleic Acids Res.*, 26:353–357, 1998.

P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov. Linguistics of nucleotide sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct.*, 6:1013–1026, 1989.

S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res*, 24(19):3836–45, 1996.

Y. Pilpel, P. Sudarsanam, and G. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–9, 2001.

K. Pruitt, T. Tatusova, and D. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33 Database Issue:D501–4, 2005.

R. Pudimat, E. G. Schukat-Talamazzini, and R. Backofen. Feature based representation and detection of transcription factor binding sites. In *German Conference on Bioinformatics*, pages 43–52, 2004.

K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23(23):4878–84, 1995.

B. Ren, F. Robert, J. Wyrick, O. Aparicio, E. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. Volkert, C. Wilson, S. Bell, and R. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–9, 2000.

A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W. Marshall, and A. Khvorova. Rational siRNA design for RNA interference. *Nat Biotechnol*, 22(3):326–30, 2004.

F. R. Roth, J. D. Hughes, P. E. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, 16:939–945, 1998.

E. Roulet, S. Busso, A. Camargo, A. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20(8):831–5, 2002.

P. Saetrom. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics*, 20 (17):3055–63, 2004.

P. Saetrom and O. Snove, Jr. A comparison of siRNA efficacy predictors. *Biochem Biophys Res Commun*, 321(1):247–53, 2004.

A. Sandelin and W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2):207–15, 2004.

A. Sandelin, W. Alkema, P. Engstrom, W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32 Database issue:D91–4, 2004a.

A. Sandelin, W. W. Wasserman, and B. Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):249–252, Jul 2004b.

M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *J. Mol. Biol.*, 297:599–606, 2000.

U. Schibler. Circadian rhythms. Liver regeneration clocks on. *Science*, 302 (5643):234–5, 2003.

T. Schneider and R. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, 1990.

T. Schneider, G. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188(3):415–31, 1986.

S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res.*, 10(4):577–586, 2000.

S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proc. int. conf. intell. syst. mol. biol.*, volume 8, pages 344–54, Menlo Park, California, 2000. AAAI Press.

G. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res*, 30(1):328–31, 2002.

B. Swanson, H. Jack, and G. Lyons. Characterization of myocyte enhancer factor 2 (MEF2) expression in B and T cells: MEF2C is a B cell-restricted transcription factor in lymphocytes. *Mol Immunol*, 35(8):445–58, 1998.

M. Truss, M. Swat, S. Kielbasa, R. Schafer, H. Herzel, and C. Hagemeier. HuSiDa–the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells. *Nucleic Acids Res*, 33 Database Issue:D108–11, 2005.

K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki, A. Juni, R. Ueda, and K. Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res*, 32(3):936–48, 2004.

A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62, 2003.

D. Vallone, S. Gondi, D. Whitmore, and N. Foulkes. E-box function in a period gene repressed by light. *Proc Natl Acad Sci U S A*, 101(12):4106–11, 2004.

J. van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

J. van Helden, M. del Olmo, and J. E. Perez-Ortin. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, 28:1000–1010, 2000a.

J. van Helden, A. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28:1808–18, 2000b.

V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

A. Wagner. A computational genomics approach to the identification of gene networks. *Nucleic Acids Res*, 25(18):3594–3604, Sep 1997.

T. Wang and G. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–80, 2003.

W. Wasserman and J. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81, 1998.

W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87, 2004.

W. Wasserman, M. Palumbo, W. Thompson, J. Fickett, and C. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet*, 26(2):225–8, 2000.

T. Werner, S. Fessele, H. Maier, and P. Nelson. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J*, 17(10):1228–37, 2003.

D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, and L. Wagner. Database resources of the national center for biotechnology. *Nucleic Acids Res.*, 31(1):28–33, 2003.

E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, 24:238–241, 1996.

E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–9, 2000.

C.-H. Yuh, H. Bolouri, and E. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279: 1896–1902, 1998.

P. Zamore, T. Tuschl, P. Sharp, and D. Bartel. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, 101(1):25–33, 2000.

M. Q. Zhang. Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.*, 9:681–688, 1999.

J. Zhu and M. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15:607–611, 1999.

J. Zuber, O. I. Tchernitsa, B. Hinzmann, A. C. Schmitz, M. Grips, M. Hellriegel, C. Sers, A. Rosenthal, and R. Schäfer. A genome-wide survey of RAS transformation targets. *Nature Genetics*, 24:144–52, 2000.

Curriculum vitae

Personal data

Dipl.-Phys. Szymon M. Kiełbasa
Schloßstr. 68
D-12165 Berlin
Tel.: (+49) (30) 84 316 317
E-mail: s.kielbasa@itb.biologie.hu-berlin.de

Date of birth: March 12$^{\text{th}}$, 1973
Place of birth: Kraków (Cracow), Poland
Citizenship: polish

Degree

M.Sc. in Physics, Jagiellonian University, Cracow, Poland

Diploma thesis: *A strong semiconductor laser for analysis of gas properties at dielectric surfaces.*
Final exam date: Sep. 8$^{\text{th}}$, 1998
Final exam grade: very good

Education

Dec. 1999 – Dec. 2004	PhD student in the Institute for Theoretical Biology, Humboldt University, Berlin, Germany Group of Prof. Hanspeter Herzel
Oct. 1992 – Sep. 1998	Student of Physics at the Jagiellonian University, Cracow, Poland
Sep. 1987 – Jun. 1992	Second High School of King Jan III Sobieski, Cracow, Poland

Stipendiums/Fellowships

Apr. 2002 – Dec. 2004	e-fellows.net (Germany) Online-stipendium for "herausragende akademische Leistungen"
Oct. 2003 – Nov. 2003	Boston University, Boston (USA), cooperation (5 weeks) with the group of Prof. Zhiping Weng
Oct. 2002 – Nov. 2002	Boston University, Boston (USA), cooperation (5 weeks) with the group of Prof. Zhiping Weng
Jul. 2001	Student travel fellowship, 9th International Conference on Intelligent Systems for Molecular Biology (Copenhagen, Denmark)
Apr. 1996	Rotary Foundation: 5 weeks fellowship, the hardware development department of IBM (Vimercate Plant, Italy)

List of publications

S. Kielbasa, D. Gonze, and H. Herzel. Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6:237, 2005

M. Truss, M. Swat, S. Kielbasa, R. Schafer, H. Herzel, and C. Hagemeier. HuSiDa–the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells. *Nucleic Acids Res*, 33 Database Issue: D108–11, 2005

N. Blüthgen, S. Kielbasa, and H. Herzel. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res*, 33(1):272–9, 2005b

S. Kielbasa, N. Blüthgen, and H. Herzel. Genome-wide Analysis of Functions Regulated by Sets of Transcription Factors. *Proceedings of the German Conference on Bioinformatics*, pages 105–113, 2004a

S. Kielbasa, N. Blüthgen, C. Sers, R. Schäfer, and H. Herzel. Prediction of Cis-Regulatory Elements of Coregulated Genes. *Genome Informatics*, 15(1):117–124, 2004b

N. Blüthgen, S. Kielbasa, B. Cajavec, and H. Herzel. HOMGL-comparing genelists across species and with different accession numbers. *Bioinformatics*, 20(1):125–6, 2004

Z. Hu, Y. Fu, A. Halees, S. Kielbasa, and Z. Weng. SeqVISTA: a new module of integrated computational tools for studying transcriptional regulation. *Nucleic Acids Res*, 32(Web Server issue):W235–41, 2004

S. Kielbasa, J. Korbel, D. Beule, J. Schuchhardt, and H. Herzel. Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*, 17(11):1019–26, 2001

H. Herzel, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, and J. Schuchhardt. Extracting information from cDNA arrays. *Chaos*, 11(1):98–107, 2001

Acknowledgements

First of all, it is a pleasure to thank Prof. Hanspeter Herzel for giving me the opportunity to work on my Ph.D. at the Institute for Theoretical Biology (ITB, Humboldt University, Berlin). In particular, I wish to thank him for the careful and patient supervision of my work, numerous valuable discussions and excellent working conditions. I feel deeply grateful and privileged to have been his student.

Furthermore, I would like to express my gratitude to Prof. Herzel, Prof. Reinhart Heinrich and the Graduate Program 268 (supported by German Research Foundation) for providing me a chance to participate and present my results at several international conferences all over the world. Here, I thank also Prof. Zhiping Weng for accepting me in her lab and giving me a chance to collaborate with her collegues Dr. Martin Frith and Dr. Zhenjun Hu.

I am deeply indebted to Maciej Swat, Nils Blüthgen, Dr. Didier Gonze, Branka Čajavec, Dr. Dieter Beule, Dr. Johannes Schuchhardt, Stefan Legewie, Jan Korbel and all other colleagues of the group of Prof. Herzel for their help and the great ambience. We had a lot of lively discussions about the research and they gave a lot of valuable comments, which significantly improved the papers we wrote and this thesis as well.

The German Federal Ministry of Education and Research (BMBF) and the German Research Foundation (DFG) are acknowledged for the financial support.

Last but not least, I would like to thank my family, who enabled my studies, and Ewa who recently gave me the necessary support.

Selbständigkeitserklärung

Hiermit erkläre ich, dass alle verwendeten Hilfsmittel und Hilfen in dieser Arbeit angegeben worden sind. Ich versichere, dass die Dissertation auf dieser Grundlage von mir selbständig erarbeitet und verfasst wurde.

Berlin, 12. März 2005

Szymon M. Kielbasa